CREDIBILITY ASSESSMENT FOR ARABIC MICRO-BLOGS USING NOISY LABELS

Almansour, Amal

Awarding institution:
King's College London

CREDIBILITY ASSESSMENT FOR ARABIC MICRO-BLOGS USING NOISY LABELS

By Amal Abdullah AlMansour

A thesis submitted in partial fulfilment of the requirement for the degree of Doctor of Philosophy

Department of Informatics

Kings College London, University of London

2016

Abstract

Due to their openness and low publishing barrier nature, User-Generated Content (UGC) platforms facilitate the creation of huge amounts of data, containing a substantial quantity of inaccurate content. The presence of misleading, questionable and inaccurate content may have detrimental effects on people's beliefs and decision-making and may create a public disturbance. Consequently, there is significant need to evaluate information coming from UGC platforms to differentiate credible information from misinformation and rumours. In this thesis, we present the need for research about online Arabic information credibility and argue that by extending the existing automated credibility assessment approaches to adding an extra step to evaluate labellers will lead to a more robust dataset for building the credibility classification model.

This research focuses on modelling the credibility of Arabic information in the presence of disagreed judging credibility scores and ground truth of credibility information is not absolute. First, in order to achieve the stated goal, this study employs the idea of crowdsourcing whereby users can explicitly express their opinions about the credibility of a set of tweet messages. This information coupled with the data about tweets' features enables us to identify messages' prominent features with the highest usage in determining information credibility levels. Then experiments based on both statistical analysis using features' distributions and machine learning methods are performed to predict and classify messages' credibility levels. A novel credibility assessment model which integrates the labellers' reliability weights is proposed when deriving the credibility labels for the messages in the training and testing dataset. This credibility model primarily uses similarity and accuracy rating measurements for evaluating the weighting of labellers.

In order to evaluate proposed model, we compare the labelling obtained from the expert labellers with those from the weighted crowd labellers. Empirical evidence proposed that the credibility model is superior to the commonly used majority voting baseline compared to the experts' rating evaluations. The observed experimental results exhibit a reduction of the effect of unreliable labellers' credibility judgments and a moderate enhancement of the credibility classification results.

Table of Contents

Table of Figures

Table of Tables

Acknowledgements

In the name of Allah, Most Gracious, Most Merciful. All thanks and praise are due to Allah and peace and blessings be upon his messenger. Writing this thesis has been a journey of learning, from the initial concept to the final product, and through this journey I would like to thank some of the people who helped me attain this goal. First of all I would like to thank Professor Costas Iliopoulos, my supervisor, who was always approachable and helpful. I thank him for his continued support, constructive comments and encouragement. Without his excellent guidance, this work would not have taken the present shape. I also want to thank Professor Ljiljana Brankovic from the University of Newcastle - Australia for her guidance and constructive comments at different stages of my study. In addition, reaching my goal has a lot to do with my friends and PhD colleagues who helped me by constantly providing support and encouragement. Finally, I would express a deep sense of gratitude to my parents, my sisters and brothers for their understanding, patience and love. Without their prayers and support, I would not have been able to make it this far.

I would also like to acknowledge the funding support from King Abdul-Aziz University (KAU) - Saudi Arabia; I would like to thank them for the scholarship and the continuous support throughout the years of my study.

Abbreviations

Abbreviation	Meaning
UGC	User Generated Content
NLP	Natural Language Processing
NDCG	Normalized Discounted cumulative gain
CIT	Conditional Independence Test
SVM	Support Vector Machine classifier
PRF	Pseudo Relevance Feedback
TF-IDF	Term Frequency–Inverse Document Frequency
IPIP	International Personality Item Pool
PCC	Pearson's Correlation Coefficient
ICC	Intra-class Correlation Coefficient
LIWC	Linguistic Inquiry and Word Count

1 Introduction

Web 2.0 is the term used to describe World Wide Web sites that utilize User-Generated Content (UGC) concept. The main idea behind these sites is to allow their members to create and share online information freely and easily as it does not require any web design or publishing skills. There are a number of different types of Web 2.0 platforms including wikis, blogs, forums, social networking, video/photos sharing sites, question-answer sites, recommendation sites, and product reviews. In these platforms, when users are the role players; they are both the information consumers as well as the information producers; in contrast to traditional Web sites where individuals are passive viewers of content.

Recently, UGC platforms are widely used as primary vehicles for sharing and spreading news, information, opinions and experiences between people around the world and will continue to grow in popularity. For example, articles about a variety of topics are available on wikis and blogs, while advice is being shared on question-answer sites and social networks. Furthermore, opinions are freely exchanged through blogs, microblogs and social networks [1]. Research by Flanagin and Metzger 2000 [2] revealed that individuals depend more on online sources for seeking information and it is fast replacing traditional sources, such as books, newspapers, and television. Indeed, social networking sites allow their members to easily and freely disseminate "real-time" news updates to a large number of people - in some cases, before traditional media [3]. According to a report published by Pew Research Centre on October 5, 2012, declared that the percentage of Americans who checked news on a social networking site has doubled from 9% to 19% – since 2010[1].

However, On the bright side, these platforms opened unprecedented communication possibilities with nil or minimal control over information flow as they enable its members to easily and freely publish whatever they like and make it widely accessible. On the downside, this low publishing barrier can lead to the creation of huge amounts of data, containing a substantial quantity of inaccurate and misleading content. Most of the published information in UGC platforms lack any type of quality control and may not face the same scrutiny and review processes as other conventional mass media. This issue becomes problematic as more people search and browse UGC platforms for sensitive topics regarding health, politics, business, and crises news. Since our society, especially the younger generation [4], might be influenced by

[1] http://www.pewresearch.org/daily-number/social-networkingsites-grow-as-a-source-of-news/

information from UGC platforms, the presence of misleading, questionable and inaccurate information may have detrimental effects on their beliefs, judgements and unnecessarily provoke panic and may cause a undue alarm or knee-jerk responses. Consequently, there is significant need to evaluate information generated from UGC platforms to differentiate credible information from misinformation and rumours.

1.1 Motivation and Research Objectives

Credibility has been studied in different UGC platforms, with particular emphasis on the micro-blogging service Twitter, which acts not only as a social network, but also as a news source [5], [6]. On the global front, Twitter has become a popular social medium for sharing real-time information. Indeed, by allowing its members to easily and freely publish whatever they like and enabling transmission of news have also marked it as a potential place for rumours and gossips. Therefore, there is urgent demand for models that can assess the credibility of tweet messages as it is a popular source for real-time information seekers around the world. In this research, different automatic credibility assessment methods related to the micro-blogging platform-Twitter have been examined and evaluated for the purpose of extending research on assessing online information credibility. Based on our survey study [7], we identified several concerns can be tracked in this field within the existing models discussed in literature review. The existing literature highlights some key limitations associated with assessing the credibility of tweet messages and tackling these issues serves as the motivation of this study. Key limitations are outlined below followed by more details for each one:

1. **Limited use of credibility assessment models in Arabic context:** This study describes the need to evaluate Arabic information credibility in UGC (Twitter as a test case).
2. **Reliance on crowd labellers' ratings to obtain the ground truth for tweet messages' credibility:**
    a. This study describes the problem of having noisy labelled dataset with increased chance for labellers' disagreements provided by a set of crowd with varying reliability levels.
    b. It also describes the problem of achieving objective credibility scores of tweet messages from multiple subjective judgments obtained from semi-anonymous labellers with un-known expertise.

3. **Limited evaluation of Arabic credibility prominent features using both explicit and implicit methods:** This study describes the need to combine both explicit and implicit methods to identify Arabic credibility prominent features. Explicit method using users' feedback and implicit method using the analysis of tweets' features distributions.

1.1.1 Credibility in Arabic context

With the growing popularity of UGC platforms as a communication medium for sharing and disseminating information among people around the world, detecting message credibility across different languages is imperative. According to our survey study on previous credibility evaluation efforts [7], most prior credibility models were based on English content and labelled by mainly English speaking users with a few in Spanish and Chinese. In terms of the Arabic language, there have been very few efforts [8], [9], although it is clear the role of social media, in particular Twitter, in the political changes of 2010-11 (Arab Spring) and how the social media in the Middle East is growing rapidly particularly amongst young people. Also it should be noted that there is a lack of existing Twitter datasets regarding Arabic language which means there is a need to create a publicly available labelled dataset covering different topics that can be used for further analysis. Indeed, as noted from the survey study [7], credibility perception is subjective: different community groups may have different opinions, attitudes, and preferences while consuming UGC information messages thus might be perceived according to different credibility levels. Therefore, credibility assessments need to be considered relative to both people credibility judgments and credibility contexts such as environment, situations, expectations, etc. [10], [11]. Within this framework, exploring credibility assessment in other culture and language such as Arabic would be a compelling area of research.

Research Objective#1:

Needless to say, automatic credibility assessment in Arabic settings has not yet been thoroughly targeted by the research community, since the majority of proposed models are only tested on English datasets; not much research has been conducted to assess credibility of Arabic content. Consequently there is a need to apply previous assessment methods and investigate their usefulness with Arabic content. Also, it was noticeable that content features had a profound impact on detecting credibility in previous studies. It is recommended to identify the content features that had been previously studied in English content and investigate if some

of them could be substituted with other features relative to Arabic content (case sensitivity of English words).Therefore, the first objective for this research study is to evaluate Arabic content credibility in UGC (Twitter as a test case) in bridging the gap in Arabic credibility research, while extending the current web credibility research across language families. This study provide an analysis regarding credibility in the context of three use cases: 1) using Arabic language 2) covering various topics and 3) labelling associated with Arab viewers. One of the outcomes of the presented study, is a corpus of human annotated Arabic messages along with computed features where some are novels that could be used for further research.

1.1.2 Constructing credibility ground truth

Supervised learning-based approaches have been used prominently in the field of credibility classification as noted from the survey study [7]. These approaches require building a training/testing dataset that typically contains a collection of tweet messages together with optional features. The messages in these datasets are then labelled by human labellers which range from crowd sourcing, via volunteers, to experts in the topic field. Then, the collected labels are aggregated to estimate the true labels. Human labellers' judgments of messages' credibility is important step for evaluating the automatic credibility models and can also be used as a source of training data for machine learning methods to build their models [8], [12]–[18]. It is well known that having a high quality dataset, and in particular a correct credibility assessment, is a necessary condition for building a successful automatic credibility assessment model, and correctly evaluating its predictive accuracy. Therefore, human labellers can be seen as an important factor affecting the prediction model [19].

Most previous studies rely on human labellers from crowdsourcing websites; not trusted media sources or experts in topic domain; to obtain their ground truth labelling for credibility classification as expert labellers' judgments for corpus labelling in practice are difficult, and expensive to obtain. Also, there are no studies to investigate the option to estimate objective labels considering labellers' disagreement. Indeed, many labelling tasks related to opinion are subjective by nature as the labelling in this research and thus there is no clear correct label. Most of previous studies which encounter messages with conflicting credibility scores resolved them by an extra judgment [12], labelling them "unsure" [13], or just discarding them [14]. In addition, most surveyed research did not pay sufficient attention to analyse labellers' reliability to justify the quality of their credibility ratings. Even though, these online labellers are generally

semi-anonymous and unverified, hence it is difficult to ascertain about their expertise and reliability hence the quality of their ratings. In this thesis, we argue that there are differences between labellers' reliability levels that might affect their credibility judgments and ought to be taken into account. Therefore, evaluating labellers' reliability through examining their credibility score ratings is important for obtaining a correct credibility labels and considered a noteworthy research direction.

Research Objective#2:

A crucial step to building a credibility classification system is the corpus labelling where multiple subjective (possibly noisy) labels are collected for the same tweet message from crowd labellers. Even though, we assume that unpaid labelling tasks originate from volunteers; as the case in this research; is not targeted by malicious labellers, labelling tasks that are open to the public are still subject to low-quality noisy labelling, either through recklessness, laziness, or misunderstanding. In this research, an assumption of unequally reliable labellers is proposed and that the reliability of each labeller is unknown, as is the correct credibility of the tweet messages.

Research aim is to infer the correct credibility labels of the tweet messages by estimating the reliability of the labellers. To achieve high-quality labelling hence building more robust classification system, it is necessary for additional examination against crowd labellers' annotation and to extract as much information from labellers' judgments as possible before it is combined to estimate the correct labels. A mechanism to evaluate labellers' reliability is desirable to reduce the influence of unreliable crowd labellers and to produce more objective ground truth labels taking into account labellers' disagreement. We propose a theoretical credibility assessment model that takes labellers' reliability differences into account when estimating the correct messages' credibility labels. This model uses different criteria measurements for evaluating the reliability of labellers in purpose to maximize the weight of high reliable labellers and reduce the influence of unreliable crowd labellers.

1.1.3 Twitter Arabic content prominent features

Even though previous research already proposed feature-driven approaches to assess credibility, it did not investigate the usefulness of these features in informing credibility judgments in different contexts such as different culture, language, topic and situation. Also

incorporating user surveys' results in identifying the credibility features importance has not been taking lots of attention in previous work. In addition, there is variation in how credibility surveys are conducted considering the number of features presented to the participants and the studied effect. Most of the previous surveys mainly manipulate data (i.e., user images, user names, etc.) within their experiment to measure its impact on users' credibility judgments. However, to identify the effectiveness of various features for information credibility classification in Arabic context, this study suggests presenting all tweet messages along with their cues/features to labellers. Then labellers are requested to rate the credibility of messages and also a set of Twitter content and author features.

Research Objective#3:

In this research, we aim to identify messages' prominent features that have most usage in determining information credibility levels, by using credibility ground truth corpus of human annotated Arabic messages that has been built for this purpose. By identifying these features, we enrich existing automatic credibility classification techniques, for example by adjusting specific features' weights depending on their occurrence and importance to the end users. We suggest using implicit and explicit methods to check the prominent features consumed to assess the tweet messages credibility. For the implicit method, a histogram is used to detect the percentage of occurrences of these features in different credibility classes, where for the explicit method, it would be a good idea to incorporate feedback from users surveys designed to rate the importance of features on assessing messages' credibility. In terms of user survey, we will present a list of prominent features from previous studies and advise users to select to what degree this feature convey tweet credibility.

Based on previous work, we distinguish three main groups of features: authority and topical expertise (of the source), data quality (of the content), and popularity (of the content and the source). This study would like to investigate the effectiveness of these groups of features in determining credibility in Arabic context using the following assumptions:

1. Having a source with authority and topical expertise will increase the credibility of tweets. (Larger friends count and followers' count, statuses count showing source experience in using microblogging and being an active, old registration age, presence of description with extra self-disclosure degree, verified account).

2. Having a better data quality of tweet content will increase the credibility of tweets. (Longer tweets, inclusion of URL, Less number of !/?, No inclusion of person pronoun, No unique characters, No swear words, No presence of sad / happy emoticons, inclusion of number of hashtags).
3. Having a tweet message with higher popularity will increase the credibility of tweets. (Higher retweet count, number of hashtags).
4. Having a source with higher popularity will positively affect credibility of tweets. (Larger followers count, statuses count, old registration age).

1.2 Research Questions

Reviewing previous credibility studies provoke the following research questions that would help to illustrate the steps towards realising this thesis. Here are a few questions, and pointers related to my research study:

- Is multiple labelling for the same tweet messages by different labellers yield same judgments or it may contain disagreed and noisy labels? How these noisy subjective labels can best be used to infer the correct objective labels of the tweet messages?
- How good and reliable are the credibility assessments of volunteered labellers? How can we estimate the reliability of the labellers? What methods should be used to quantify the reliability?
- Is there any relation between labellers' data including: demographics, Twitter usage, and topics' familiarity and the perception of credibility?
- Would identify reliable labellers and weighting them higher built a better system to predict credibility and improve the accuracy for the classifier model? Is having a higher majority agreement levels positively impact the learning performance and accuracy of machine learning algorithms?
- Do labellers with similar credibility ratings have similar credibility features??
- What type of available features is most useful for informing credibility judgment's about Twitter information in Arabic contexts?

1.3 Research Main Contributions

To evaluate these questions, the unique main contributions to the information credibility field made in this research are outlined below:

- Offer a systematic review of the current developments in assessing information credibility automatically in UGC platforms, focusing on microblogging service.
    - Present comparative analysis of credibility assessment models based on different criteria that incorporate labellers, language and culture context, used classification techniques, used features and performance evaluation.
    - Classify previously Twitter credibility surveys based on the features they considered to draw a clear view of the techniques and surveys used for micro-blogging credibility.
- Propose to assess information credibility using Arabic context; used model is developed to detect credibility in the context of three use cases: using Arabic language, covering different topics, and labelling by Arab viewers. The following contributions have been suggested for this step:
    - Report on survey data that show credibility perceptions among Arab audiences, and discuss how Arab users consume micro-blog content and how to incorporate these findings into classifying Twitter information automatically.
    - Identify messages' features and factors that demonstrate user preferences within Arabic settings.
- Investigate inter-labeller agreement values in different settings: different labellers, different presentations, different topic types, and different number of labellers.
- Propose a novel information credibility assessment model which attempts to infer the correct tweet messages labels given noisy subjective labels. The model determines the true credibility labels of the tweet messages by estimating the reliability of the labellers. The following contributions have been recommended for this step:
    - Propose number of measurements to evaluate labellers' reliability and validate how fair and reliable their credibility scores in labelling the tweet messages corpus. We proposed different measures to identify both: the pairwise relation between labellers rating values and between the labellers' rating values and the average rating values.

- Use iterative algorithms based on selected measurements for updating labellers' reliability weights and evaluate the stability of their weights through iterations.
- Construct credibility ground truth values taken into account labellers' reliability weights.

- Use machine learning methods, mainly supervised learning techniques, to study how, and to what extent, the existing supervised learning algorithms can be used to identify credible information in Arabic context.
- Explore the relation between the level of majority agreement and machine learning performance by comparing agreement level to classification results. Ascertain if labellers conclude almost the same judgemental labelling hence machine learning algorithm would outperform, resulting in better classification results.

1.4 Research Methodology

The increased usage of Twitter as a medium for reporting news and sharing information between people has caught the attention of researchers from different disciplines. One of the research directions is the analysis of online information from the perspective of its credibility. This research aims to assess and analyse the credibility of tweet messages in Arabic language. In order to achieve the research stated goals, we will undertake the following three main phases:

First Phase: involves employing the idea of crowdsourcing where users can explicitly express their opinions about credibility of a set of tweets. This information coupled with the data about tweets' features enable us to investigate which features may indicate the credibility level of a tweet, e.g. tweet with attached image and was authored by a person who posts a lot of tweets will be, with high probability, a credible tweet. In addition to this, basic analysis of credibility rating values and labellers' data characteristic has been conducted to examine the Arabs perception of credibility followed by a study of inter-labellers agreement using different settings. Through the agreement study, we identified three experts who also rated the credibility of tweets and based on that we investigate the level of agreement between experts and the crowd, and we identify which expert represents the crowd in the best way. This can allow us to select the most representative expert when it is needed.

Second Phase: it is the main part of the study where we apply our proposed model based on the following steps: extracting labellers' traits and credibility rating scores, evaluating labellers'

reliability, and finally constructing credibility labelling. In the first step, basic traits of each labeller are extracted explicitly from the user survey along with his/her labelling scores. In the second step, Labeller's credibility scores for the tweet messages are used as inputs to generate labellers' reliability weights using mainly accuracy and similarity measurements. The last step uses labellers' reliability weights to construct the correct credibility labels for the tweet messages. In this proposed framework, we applied different measurements to weight the labellers and used experiments to assess how the proposed techniques can enhance the quality of the applied dataset and reduced the spontaneity of judgments. In order to evaluate proposed model, we compare the labelling obtained from the expert labellers and those from the non-expert labellers, and expect their ratings values will exhibit a superior credibility judgment similar to experts.

Third Phase: it is the last phase of the study where we use the constructed labels with feature-based approaches mainly: relative features frequency model and decision tree algorithm to detect credibility of tweet messages. In this phase, features related to tweets are computed and analysed. Implicit and explicit methods have been used to check the prominent features consumed to assess the tweet messages credibility. For the implicit method, a histogram was used to detect the percentage of occurrences of these features in different credibility classes, where for the explicit method, a user survey was used to rate the importance of features on assessing messages' credibility.

1.4.1 Proposed system structure

While reviewing the literature, we identified four essential stages for building feature-driven credibility classification model: 1) Dataset collection, 2) Labelling, 3) Feature extraction and analysis, and 4) Classification. In this research, we introduce an extra step prior to the classifier building to evaluate labellers' credibility judgements which solve the problem of labelling disagreements and produce more objective tweet messages labels. The proposed system architecture is illustrated in Figure 1-1.

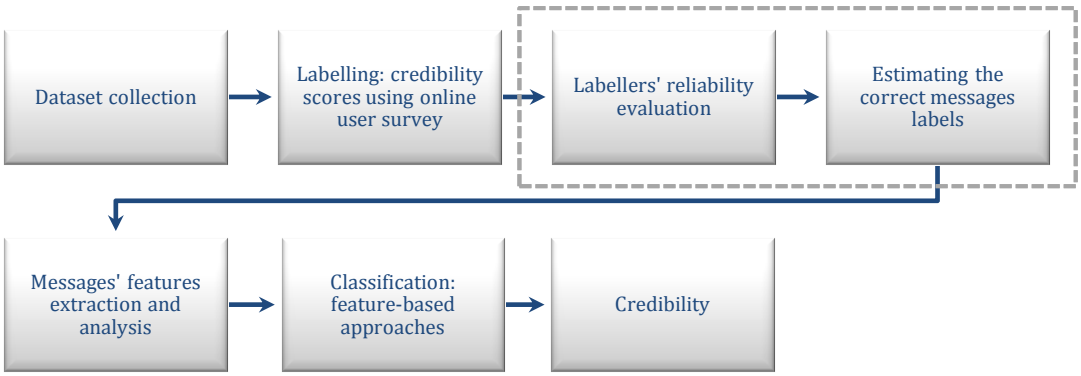


Figure 1-1 Proposed system architecture

1.5 Research Significance

As more individuals construct their opinions and actions based on online content, the unfiltered and distributed nature of social networks might lead to spreading baseless rumours amidst valid truthful news. Potentially dangerous consequences may result from relying on incredible content - a research by Jensen et al. 2011 [20] stated that there are numerous cases show how deceptive information can severely harm business and society. Below are some examples of how consuming online information especially from UGC content might affect one's decision-making ability and possibly adversely impact our society.

Business - Blogs: Apple shares[2]**:** The fallout resulting in the decline of Apple shares in October 2008 after a blog rumour which stated that the founder and CEO Steve Jobs had suffered a heart attack is an example of how false reports may cause unnecessary distress.

Crises - Twitter: Chile earthquake[3]**:** Another example regarding Twitter occurred following the Chile earthquake in 2010. Conversely during a crisis, groups also use Twitter for search and rescue missions resulting in many positive and uplifting stories which reinforces how micro-blogging can be an invaluable tool. Undeniably, there are fabricated tweets which may hamper rescue results, such as the following tweet in Figure 1-2.

Figure 1-2 Snapshot of false tweet - Chile earthquake[4]

[2] http://money.cnn.com/2008/10/03/technology/apple/
[3] http://irevolution.wordpress.com/2010/06/30/crowdsourcing-detective/
[4] http://irevolution.wordpress.com/2010/06/30/crowdsourcing-detective/

In response to this tweet, the police sent emergency personnel from the rescue department to that address, but then discovered there were no collapsed buildings and unfortunately, it was a false tweet!

Politics – Facebook and Twitter : Arab Uprising (e.g. Egypt revolution): Lynch [21] claims in his book "The Arab Uprising: The Unfinished Revolutions of the New Middle East" that social networking sites, in particular Facebook and Twitter, helped young Arab generation to share and discuss their political opinions, which helped to organize the Arab Uprising. Also it is worth mentioning that blogs, forums, and microblogs (tweets) are not only used by individuals to express their thoughts and spread information but also by news and government agencies to capture the pulses of the public at different events, such as what happened during the Arab Uprising.

Health – Websites: H1N1 Flu: In October 2009, the Food and Drug Administration FDA[5] issued notice letters to sites publishing false information about treatment products that claim to help with H1N1 influenza virus.

All examples discussed above show the dependence on the online content as source of information in many crucial domains. It also presents the detrimental effects of misleading and inaccurate information may cause to business and society. Evidently, there is an urgent need to establish online information credibility assessment, specifically in the health, crises and business sectors where the risk of low-credible content is unwieldy.

1.5.1 Why Arabic Twitter messages

Most of the research in automatic credibility assessment of Web pages and UGC platforms content has been conducted in English or Chinese, as shown by the following literature review chapter. There have been minimal efforts attributed to the Arabic language although the role of social media in the Middle East is thriving. In his article "Twitter takes off in Saudi – and other news of social media in the Arab world", August 2013, Damian Radcliffe from BBC[6] stated that there is relative popularity for Twitter among other social media platforms in the Arab region and much of the Twitter usage is dominated by users from Saudi Arabia and Egypt. In fact, the number of Saudi Twitter users doubled - up by 128%; more than a million of joined in the past

[5] http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm187142.htm
[6] http://www.bbc.co.uk/blogs/collegeofjournalism/entries/832a893c-5bf3-3ea1-89eb-79e0caf6945f

year making the number raised to 1.9 million Twitter users. Saudi Twitter users make up just over half of the Middle East's 3.7 million Twitter users, and on a daily basis produce almost half of region tweets, of which 90% are in Arabic. Figure 1-3 shows the number of Twitter users in the Arab region (Average number for March 2013).

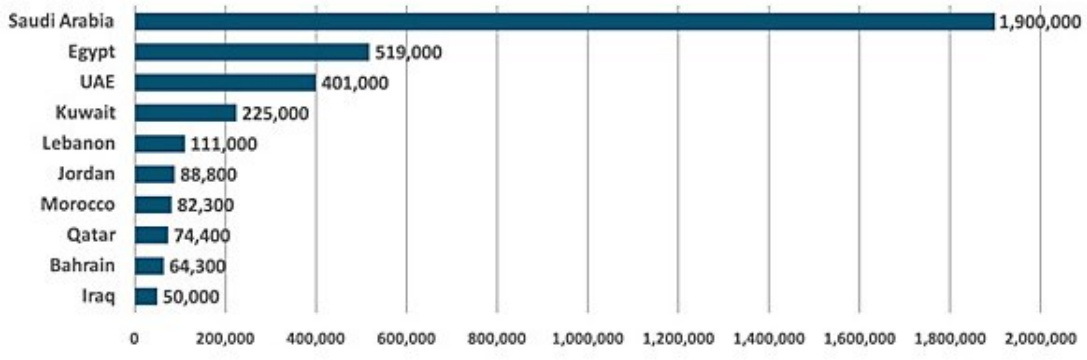


Figure 1-3 Number of Twitter users in the Arab region (Average number for March 2013)[7]

According to our survey on previous efforts, automatic credibility assessment in Arabic setting has not yet been thoroughly targeted by the research community. Therefore it is necessary to create a publicly available labelled dataset as well as to evaluate the current state of the art of research on web credibility. Among various UGC platforms, we limit our focus to Twitter, for its popularity among Arab users and for being considered the most pertinent social medium used as news source [5], [6].

1.6 Thesis Structure

This thesis consists of six chapters and has the following structure:

Chapter 1 is an introduction for the subject of the thesis; it describes the thesis background, overall thesis objectives, methods of research, the significance of the research study and the thesis organization.

Chapter 2 provides an overview to key concepts and draws a clear view of the relevant research used for assessing Twitter information credibility automatically along with some selective studies from other UGC platforms. The main objective in this chapter is to examine existing automatic information credibility classification methods based on different criteria to outline which part require more exploration.

[7] http://www.bbc.co.uk/blogs/collegeofjournalism/entries/832a893c-5bf3-3ea1-89eb-79e0caf6945f

Chapter 3 is dedicated to the employed methodology and the characteristics of the analysed dataset. The collection process and the characteristics of the newly created dataset are elaborated in this chapter in addition to a basic analysis of submitted credibility rating values and the collected labellers' data. The last sections of this chapter are devoted to the agreement calculation for the used dataset using different settings proceeded by the procedure used to construct credibility ground truth labels using majority voting method.

Chapter 4 covers the main part of the study which is dedicated to the proposed credibility model. It introduces the labellers' weighted model along with the used measures and algorithms utilized to evaluate the quality of the crowd labellers' credibility ratings.

Chapter 5 elaborates on the two important steps to detect credibility which feature extractions and credibility classification. Arabic credibility prominent features and the credibility classification accuracy results using statistical approach and machine learning classifier are discussed by this chapter.

Finally, chapter 6 presents the conclusions based on the results obtained and provides insight into solving the proposed research questions. Also it provides some recommendations for further research on thesis area.

At the end of the thesis, there is a listing of all the references used in the thesis followed by extensive appendices which include materials used in the experiments and that are referred to throughout the thesis.

2 Review of Literature

Currently a large body of literature exists about credibility evaluation for different UGC domains, with particular emphasis on the microblogging service Twitter. This chapter will provide background knowledge on the theoretical part of information credibility and define some essential and relevant terms to our research survey topic. It also draws a clear view of the relevant research used for assessing Twitter information credibility automatically along with some selective studies from other UGC platforms. Our main objective in this chapter is to examine existing automatic information credibility classification methods based on different criteria to outline which part require more exploration.

2.1 Credibility

In order to begin discussing previous studies of information credibility assessment, we need first to formulate and define some essential and relevant terms to our research survey topic:

2.1.1 Credibility definition and components

Despite the long history of credibility research that dates back to the 1951 (Hovland & Weiss 1951) [22] there is as yet no clear definition of credibility. Credibility is a multifaceted concept and has been defined as believability, trust, reliability, accuracy, fairness, objectivity, and dozens of other concepts and combination of them [23]. Simply, the overarching view across definitions, credibility has been more closely correlated to "believability" of a statement, action, or source. For the purpose of this work, we will use Rieh's 2010 [24] definition of credibility which is "people's assessment of whether information is trustworthy based on their own expertise and knowledge". There are numerous other definitions, including Oxford Dictionary[8] which defines credibility as the quality of being convincing or believable.

Fogg and colleagues [11], [25]–[27] concluded that scholars stated two main points help clarify the credibility construct. First, credibility is a "perceived quality" and when one discusses credibility, it is always from the viewpoint of the observer's perception. Second, credibility perceptions result from concurrent evaluations of multiple dimensions. Although the literature varies on how many dimensions contribute to credibility evaluations, the vast majority of researchers identify two key components of credibility: trustworthiness and expertise.

[8] http://www.oxforddictionaries.com

- **Trustworthiness** refers to the goodness or morality of the source and can be described with terms such as well-intentioned, truthful, or unbiased. A person is trustworthy for being honest, careful in choice of words, and disinclined to deceive. Information is trustworthy when it appears to be reliable, truthful, unbiased, and fair.
- **Expertise** refers to the perceived knowledge, skill, and experience of the source. It can be described with terms such as knowledgeable, reputable, or competent. Expertise is also an important key element because it is closely related to user perceptions of the ability of a source to provide information both accurate and valid. When people find that sources have expertise, they are likely to judge that information to be trustworthy. People combine assessments of both trustworthiness and expertise to arrive at a final credibility perception.

Generally, there are almost four main characteristics concerning credibility: 1) Perceptual: receiver-based, 2) Multi-dimensional: consist of multiple factors, 3) Situational/contextual: varies from one context to another and 4) Dynamic: changes over time.

2.1.2 Credibility factors

There are a wide range of factors proposed in many different credibility, trustworthiness, and quality of web information studies. In Table 2-1, we classify some of them into factors related to the main web credibility elements (author, content, and user).

Table 2-1 Credibility factors related to messages' author, content, and user

Proposed by	**Factors**			
	Author/Source	**Message Content**	**User/Reader**	**Others such as Design**
Wathen & Burkell 2002 [28] **Domain: Webpages**	Expertise, Knowledge, Competence, Trustworthiness, Credentials, and Influence.	Relevance, Currency, Accuracy, and Tailoring.	Assumptions about source or topic, Motivation (i.e., need for the information), Knowledge, Expertise regard topic, and Social location.	Surface attractiveness, Format, Design of interface, Speed of loading, Usability, Accessibility, Interactivity, and Flexibility.
Yamamoto & Tanaka 2011 [29] **Domain: Webpages**	Authority.	Accuracy (PageRank), Typicality, Currency, Coverage, Social bookmarks.	NA	NA
Kwon, ,Cho, and Park 2009 [30] **Domain: Movie Recommendation Website**	Expertise, Competence, Trustworthiness, and Similarity.	NA	NA	NA

Rieh & Belkin 1998 [31] **Domain: Webpages**	Institutional level (URL, types and name of institution), Or individual level (identification of author's affiliation and name).	Currency, Accuracy, Useful, Specificity, and Size of document.	NA	Format (graphical images, information structure), Presentation (writing style), and Speed of loading.
Fogg & Tseng 1999; Fogg et al. 2001; Fogg 2003; Fogg 2003 [11], [25]–[27] **Domain: Webpages**	Expertise and Trustworthiness	NA	NA	Real-world feel, Ease of use, Message tailoring, Commercial implications, and Amateurism.
Gil & Artz. 2007 [32] **Domain: Webpages**	Authority, Provenance, and Incentive.	Topic, Popularity, Recommendation, Related Resources, Bias, Agreement, Specificity, Age, Deception, and Recency	Context and criticality, Direct experience, User expertise, and Likelihood.	Appearance, and Limited resources
Flanagin and Metzger 2000 [2] **Domain: Webpages**	Authority (identification, contact Information, qualifications, objectives).	information Type, Accuracy, Objectivity, Currency, Coverage, Recommendation	Internet Experience (WWW use, experience, expertise, familiarity, and access) and Demographics.	External sources to validate Information.
Rubin &Liddy 2006 [33] **Domain: Weblogs**	Expertise, Identity Disclosure, Trustworthiness, Value System.	Information Quality (completeness, accuracy, appropriateness, timeliness, organization), and Literary appeal (i.e., writing style).	Match to prior expectations, Match to information need, Memory trigger (i.e., shared experiences), and Personal connection (e.g., the source is an acquaintance).	Appeals and Triggers of a Personal Nature (Design layout, Typography, Color schemes) and Curiosity trigger.
Kakol et al. 2013 [34] **Domain: Weblogs**	NA	NA	socio-economic status, Internet efficacy, and psychological traits	NA
Fogg et al. 2001 [35] **Domain: Weblogs**	NA	NA	age, gender, country of origin, and education and income levels	NA

Based on the analysis of the credibility factors reported in previous Web credibility research and listed in the table above, we found that there are some overlapping factors which can group into four main factors to make it easier to be used across different platforms and be applicable for automatic analysis. The suggested factors are as follows: authority and topical expertise (of the source), data quality (of the content), and popularity (of the content and the source). Also, design of the website is considered an important element in different studies; however, since we are planning to study the Twitter information credibility, this element is irrelevant.

2.1.3 Credibility assessment components

Assessing information credibility is a challenging task since “credibility” is a complex concept that is based on at least two key dimensions: trustworthiness and expertise [11], [28]. Both are judged by users consuming information, and therefore, credibility is dependent on users’ cognitive states, as users usually invoke cognitive heuristics to assess the credibility [10], [11]. In other words, credibility is determined by the “subjective judgment and assessment from the users” [36]. Also, credibility is considered situational and contextual as it varies from one context to another. An individual participant sometimes accepts certain information as credible primarily by relying on the context in which he/she encountered the information [10]. Credibility assessments need to be considered relative to both people credibility judgments and credibility contexts such as environment, situations, expectations, etc. [10], [11].

Inspired by Hilligoss & Rieh 2008 [10] framework, for the purpose of comparing various models for automatic assessment of credibility, we identified three main components affecting UGC information credibility perception: 1) contexts such as environment, topic and situation [10], [11], 2) UGC available features (cues), and 3) reader traits and cognitive heuristics such as topic knowledge and the selection of cues in making a credibility judgment. Analysing how existing models maintain and integrate these three components will provide more comprehensive understanding of information credibility perception. Figure 2-1 illustrates the main three components affecting UGC information credibility.

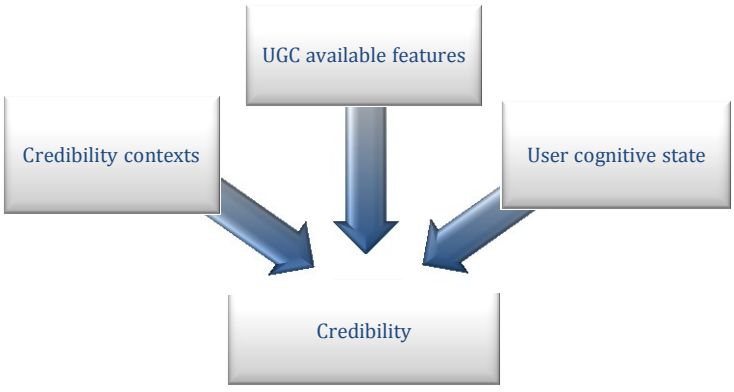


Figure 2-1 UGC information credibility components

2.1.4 Automatic assessment for information credibility

Credibility has been studied extensively in UGC domain, especially with the microblogging service Twitter, which for many people has become a popular source of up-to-date information about world events. Its ability to deliver "unmonitored" information easily and freely to users has marked it as a potential space for rumours and spams. Thus, there an urgent need to study its credibility. Most of the automated credibility assessment regarding UGC-Twitter has emerged from the fields of Data Mining, Machine Learning and Natural Language Processing (NLP). In assessing information credibility automatically, different approaches rely on different methods and factors to identify information credibility. We classified the different methods used for assessing information credibility automatically as follows: 1) Supervised learning [12]–[17], [37], [38], 2) Statistical analysis using features distribution [18] or percentage of retweets [39], 3) Similarity with credible source [8], [9], 4) Voting system based on follower relationships as vote of confidence [40], and 5) Graph analysis / Hybrid (classification with graph analysis) model [41]–[46]. More details on existing techniques are demonstrated in Table 2-2.

Table 2-2 Existing credibility assessment models

Model Used	Features Used
Supervised: Feature-based Calssification. [13], [17], [37], [38]	Different features related to the messages' author and content
Supervised: Feature-based Ranking [14]	
Supervised: Feature-based Calssification + Weighted linear combination of positive indicators. [15]	
Supervised: Feature-based Calssification + NLP. [12], [16]	Mainly linguistic features related to the messages' content
Statistical analysis : Features distributions [18]	Different features related to the messages' author and content
Statistical analysis : Percentage of retweets [39]	Propagation and manual text analysis
Content similarity with credible source: Weighted linear combination of positive indicators + NLP [8], [9]	Mainly linguistic features related to the messages' content
NLP + Voting system: Follower relationships as vote of confidence [40]	Number of followers and Topical similarity
Graph analysis / Hybrid (Classification with Graph analysis) [41]–[43], [45]	Different features related to the messages' author and content including Graph measures.

In the following subsections we discuss each method in more detail and provide a summary of the selected relevant studies in the form of an analysis table. These evaluation analyses examine the extent to which existing credibility models integrate the three main component affecting UGC information credibility based on the following criteria:

1. **Topic genera and how critical it is** (context)
2. **Dataset language** (context/ features)
3. **Level at which the credibility is assessed:** message, event (aggregated set of messages) or source credibility (features)
4. **Labelling task** is completed without prior information about the topic (believability classification) or according to given information (detecting false rumours) (user/context)
5. **Users who performed the labelling:** crowdsourcing, authors, experts or volunteers (user)
6. **Labellers' data is checked or not** (user/context)
7. **Features** describing content only (e.g., having URL), source (e.g., follower count), or aggregated set of both (e.g., fraction of tweets having URL in an event) (features)
8. **Best Predictors:** authority and expertise (source), data quality (content), or popularity (content and/or source) (features)

The first three criteria refer to the dataset and encompass context and features components, the following three refer to the labelling and encompass context and labeller components, while the last two criteria directly cover the features component. Selected studies were discussed for each model in the following subsections. The survey presented here demonstrates if these components were covered in the research of Twitter information credibility domain and to outline which part requires further exploration.

2.1.4.1 Supervised learning approach

Different supervised classifiers algorithms have been used to automatically classify or rank Twitter messages credibility. Table 2-3 categorizes previous research depending on their used algorithm and shows their classification/ranking accuracy.

Table 2-3 Predictive accuracy of supervised learning algorithms

Used by/ Algorithm	Decision Tree	SVM	Bayesian
Castillo, Mendoza, & Poblete 2011[13]	86 %		
Kang et al. 2013 [47]	58-63 %		
Kang, O'Donovan, & Höllerer 2012 [15]	88.17 %		
Gupta et al. 2013 [37]	97 %		91.9 % naive
Gupta & Kumaraguru 2012 [14]		73 % avg, NDCG Rank-SVM and PRF	
Yang et al. 2012 [38]		77.3-78.6 %	
Bhattacharya et al. 2012 [12]		73% PolyKernel and Bagging	
Qazvinian et al. 2011 [16]			93 %
Xia et al. 2012 [17]	61.2 %	66.7 %	63.6 % CIT 61.5 % K2; 62.2 % Hill Climbing

2.1.4.1.1 Decision tree algorithm

Decision tree algorithm is a very popular supervised learning method used for classification. It is based on creating a model that predicts the messages' credibility by learning simple decision rules inferred from the messages features. Tree nodes are the decision rules on one or more features and leaf nodes are predicted class labels.

Castillo, Mendoza, & Poblete 2011 [13] presented promising study for credibility classification with accuracy results 86% using J48 decision tree algorithm (Open Source Java implementation of the C4.5 decision tree algorithm [48] in Weka data mining tool [49]). They used supervised classifiers for 1) news/chat classification of 2,524 cases and 2) to assess credibility of 747 news topics. The characteristics that were taken into account included: the message itself, the user, all the tweets on the topic and the propagation of retweets. The ground truth was subjectively assessed by crowdsourcing website Amazon Mechanical Turk[9]. Their results showed that features such as propagation level, URLs inclusion, and sentiment helped effectively to classify topics automatically as credible or not credible. However, their method was based on topic credibility rather than individual tweets.

A study focused on individual tweets or users has been covered by Kang, O'Donovan, & Höllerer 2012 [15]. They identified credibility ratings of 1023 tweets collected for topic-specific (Libya). Authors trained Bayesian classifiers using manually annotated tweets, based on

[9] https://www.mturk.com

different models: 1) model uses source features, 2) model uses content features, and 3) hybrid model uses both source and content features. For the full experiment, a J48 decision tree learning algorithm was used; their best accuracy result 88.17% was obtained using the source features model.

Another study on topic-specific credibility was discussed by Kang et al. 2013 [47], who classified credible information for tweets during and after Hurricane Sandy. Annotating the two datasets was done with the assistance of Amazon Mechanical Turk users; who were asked to provide assessments, for message credibility, message newsworthiness and source credibility. Features related to message's content, author and topic were used to assess the credibility. A J48 decision tree learning algorithm has been used to predict human annotation scores based on a set of training examples. They ascertained that in both datasets: 1) message credibility, message newsworthiness and source credibility are highly correlated and 2) that both network structure and topical content of a tweet have a bearing on perceived credibility.

Research by Gupta et al. 2013 [37] studied the effectiveness of machine learning algorithms in detecting tweets containing fake image URLs in Twitter. First, they analysed the tweets for topic-specific information related to Hurricane Sandy 2012 and identified roughly 10,350 tweets containing fake images. Their analysis uncovered the following results: 1) 86% tweets spreading the fake images were retweets and 2) only a few (0.3% of the users) resulted in 90% of retweets of the fake image. Next, they analysed the overlapping between the retweets graph and the followers graph, and discovered only 11% overlap between the two graphs which means that users retweet information from other users whether they follow them or not. Finally, they used two classification algorithms: Naive Bayes and decision tree J48 to classify tweets containing fake images from real images depending on the user and tweet features. Best results were obtained from decision tree classifier, using tweet based features, which achieved 97% accuracy results. What is worth to mentioning was that their high accuracy results may be credited to the similar content of their dataset (most of the tweets were retweets).

2.1.4.1.2 SVM algorithm

Gupta & Kumaraguru 2012 [14] adopted supervised algorithms and information retrieval techniques to rank tweets based on their credibility score. Annotators were provided with news links of 14 news events to help them on labelling 500 tweets per topic. First, SVM ranking algorithm was used to rank tweets based on their content and source features. Then, the

frequent word unigrams from the top tweets were extracted and text similarity between the frequent unigrams and the top tweets was used to re-rank them using Pseudo Relevance Feedback (PRF) technique. They reached a 73% average accuracy NDCG score using Rank-SVM and PRF. Their results showed that only an average of 30% tweets contained information about the event while 14% were spam and furthermore, only 17% of the informative tweets were credible.

Estimating belief level was reported by Bhattacharya et al. 2012 [12]. They retrieved 11,591 tweets related to 32 propositions about causes and treatments which represent factual, false and debatable information. Proposition statements were identified by consulting sources such as: medical sites, physicians, and news. 2105 tweets' relevance and position (supports, opposes, other) to the probe statement were annotated with the help of oDesk crowdsourcing. When using SVM with PolyKernel and Bagging using unigram features, their accuracy results for the relevance classifier was 82% whereas the position classifier produced a 73% accuracy rate.

Rumour analysis and detection on Sina Weibo - China's leading micro-blogging service has been discussed by Yang et al. 2012 [38]. They collected a set of tweets related to rumour topics published by an official rumour busting account. Two new features were proposed, the client program and the event location in addition to the previously proposed content, user, and propagation features. The authors performed sets of experiments using SVM classifier to study the impact of incorporating these two new features in rumour classification. The classification accuracy before adding the new features were nearly 72% and when adding new features, increased to 77-78%. This study concluded that user features were more effective than content to detect rumours.

2.1.4.1.3 Bayesian algorithm

A study of rumour detection was carried out by Qazvinian et al. 2011 [16]. The authors analysed the users' believing behaviour about the rumour-related tweets and identified users that endorsed the rumour versus users who denied or questioned it. They retrieved 10,400 rumour tweets from 5 different controversial topics listed on About.com's Urban Legends site. Two annotators were asked to label each tweet if it was related to any of the rumours, or not; they use this annotation to analysis which tweets were retrieved but unrelated to the rumour. Then they asked annotators to label rumour tweet if the tweet poster believed the rumour or not;

they used the second dataset to detect users' beliefs in rumours. They built different Bayesian classifiers based mainly on the content linguistic features and then obtained a linear function of these classifiers for retrieval of the two sets. Mean Average Precision was equal to 96% in the rumour retrieval and 93% in belief classification task.

Xia et al. 2012 [17] also used a supervised method using learning CIT Bayesian Network to predict the tweets credibility in emergency situation. 350 tweets of topic-specific (England riots) were manually labelled by five experts into: credible or not credible. With the classified tweets, a number of features were extracted in relation to the user social behaviour, the content, the topic and the tweet diffusion. Classification results accuracy were between 61% using J48 and 66% using SVM, while with the proposed CIT algorithm, it reached 63%.

A summary table that examine the proposed criteria on the previous discussed studies is illustrated in Table 2-4.

Table 2-4 Evaluation of the supervised machine learning models

Criteria/Model	**Castillo et al. 2011 [13]**	**Kang et al. 2012 [15]**	**Gupta et al. 2012 [14]**	**Bhattacharya et al. 2012 [12]**	**Qazvinian et al. 2011 [16]**	**Xia et al. 2012 [17]**
Topic genera: Mixed/ Specific	M	S-Politics	M	S-Health	M	S-Politics
Language: English/ Spanish/ Arabic	**E**	**E**	**E**	**E**	**E**	**E**
Level: Message/ Event/ Source	E	M	M	M	M	M
Task: Believability classification / Detecting false rumours	B	B	B	D	D	B
Labellers: Crowdsourcing / Authors/ Experts/ Volunteers	C	—	V	E + C	V	E
Labeller data considered: Yes/ No/ Partially	**N**	**P**	**N**	**N**	**N**	**N**
Features: Content/ Source/ Group	C+S+G	C+S+G	C+S	C	C+S+G	C+S+G
Best predictors: Authority and expertise/ Data quality/ Popularity (source/content)	D+Ps,c+A	Ps,c+A	D+Ps	D	D	D+Pc,s+A

2.1.4.2 Statistical analysis approach

Mendoza et al. 2010 [39] statistically evaluated users' behaviour in crises (Chilean earthquake) with the aim to examine the ability of social network to discriminate between legitimate news and false rumours. They subsequently retrieved 42 to 700 tweets related to 7 confirmed true cases and 7 false rumours. Then, they manually labelled each tweet into: affirms

(confirming the case information), denies (refutes the case), questions (asks about the case), and unrelated. Their results confirmed that the percentage of retweets of true tweets and rumours are different and that rumours' texts were more likely to contain an indication of doubt or denial.

Research conducted by O'Donovan et al. 2012 [18] provided a statistical analysis of features distribution in four contexts: 8 different events, credible v/s non-credible messages, different length of retweet chains, and dyadic v/s non-dyadic messages - that is, tweet messages that involve conversations between pairs of users using "@" mention tag are compared to standard set of tweets. Their results in case of topics diversity and credibility levels showed that features occur more highly in topics related to crises and features such as URLs, mentions, retweet and tweet length served as good predictors for credibility. In case of retweet chains context, most notable result was the prominence of the URL feature in the longer chains, occurring in 50% of the long chain context, indicating that tweets with provenance links to other information tend to get propagated more frequently. Similarly, longer tweets in terms of words and characters tend to appear more often in longer chains. With regard to dyadic context, results showed that dyadic pairs tend to have more words, but shorter words than standard tweets. It is notable from their graph results that there is a high variance of feature occurrence across different topics which is an interesting insight to study how credibility features is distributed across different topics genera. Table 2-5 summarizes this method using the same evaluation criteria.

Table 2-5 Evaluation of the statistical analysis models

Criteria/Model	**Mendoza et al. 2010 [39]**	**O'Donovan et al. 2012 [18]**
Topic genera: Mixed/ **S**pecific	S-Crises	S-Politics
Language: English/ **S**panish/ **A**rabic	**S**	**E**
Level: Message/ **E**vent/ **S**ource	E	M
Task: Believability classification/ **D**etecting false rumours	D	B
Labellers: Crowdsourcing / **A**uthors/ **E**xperts/ **V**olunteers	A	C
Labeller data considered: Yes/ **N**o/ **P**artially	**N**	**P**
Features: Content/ **S**ource/ **G**roup	C	C+S+G
Best predictors: Authority and expertise/ **D**ata quality/ **P**opularity (**s**ource/**c**ontent)	D+Pc	D+Pc

2.1.4.3 Similarity with credible source approach

Credibility assessment has been studied for Arabic by Al-Eidan, Al-Khalifa, & Al-Salman 2010 [9] and Al-Khalifa & Al-Eidan 2011 [8] using an evidence-based method. They integrated

two approaches to evaluate the message credibility levels (low, high, and questionable). The first approach was based on computed similarity thresholds between the content of both Twitter posts and verified news sources such as SPA, Aljazeera, and Google News, where the second approach was based on a liner combination of the similarity value in addition to a set of features related to the content and the source. They evaluated their classification result against three political experts' evaluation using a dataset of 29 tweets and four news articles of two topics: (Iran) and (Yeman&Houthi). Their results indicated that the first approach was more effective in evaluating the credibility of tweets. Yet, with this approach, the system was able to assign tweets to only two credibility levels: low and high, while the second approach was able to assign the tweets to all three levels of credibility. Furthermore, linking source degree assigned by expert was the prominent feature in the second approach. It should be noted that the above method is only useful for tweets combined with credible external sources and also it did not embrace most of prominent features proposed by previous research such as hash-tags, retweets, and emoticons. An evaluation for this method is illustrated in Table 2-6.

Table 2-6 Evaluation of similarity with other source model

Criteria/Model	**Al-Eidan et al. 2010 [9] and Al-Khalifa et al. 2011 [8]**
Topic genera: **M**ixed/ **S**pecific	S-Politics
Language: **E**nglish/ **S**panish/ **A**rabic	**A**
Level: **M**essage/ **E**vent/ **S**ource	M
Task: **B**elievability classification/ **D**etecting false rumours	B
Labellers: **C**rowdsourcing/ **A**uthors/ **E**xperts/ **V**olunteers	E
Labeller data considered: **Y**es/ **N**o/ **P**artially	**N**
Features: **C**ontent/ **S**ource/ **G**roup	C+S
Best predictors: **A**uthority and expertise/ **D**ata quality/ **P**opularity (**s**ource/**c**ontent)	D

2.1.4.4 Voting approach

A work focuses on predicting experts using voting model is covered by Canini, Suh, & Pirolli 2011 [40]. They considered follower relationships as vote of confidence; they designed an algorithm using both: topic modelling analysis and social status of users to generate a ranked list of relevant and credible users for any given topic. Firstly, the algorithm use Twitter search to identify users (voters) who are associated with a query topic. Next, it filters and ranks the results by identifying users (candidates) whose followers appear frequently in the search result. Finally, they use topic modelling to analyse the textual content of the highest-scoring users and re-rank

them by this criterion. To evaluate their algorithm, they performed five search queries using Amazon Mechanical Turk participants. Comparing algorithm rankings with rankings provided by WeFollow website, their algorithm showed enormous potential to help users identify interesting users to follow in Twitter. An evaluation for this method is displayed in Table 2-7 using the same evaluation criteria in the previous method.

Table 2-7 Evaluation of the voting model

Criteria/Model	Canini et al.2011 [40]
Topic genera: **M**ixed/ **S**pecific	M
Language: **E**nglish/ **S**panish/ **A**rabic	**E**
Level: **M**essage/**E**vent/ **S**ource	S
Task: **B**elievability classification / **D**etecting false rumours	—
Labellers: **C**rowdsourcing / **A**uthors/ **E**xperts/ **V**olunteers	V+C
Labeller data considered: **Y**es/ **N**o/ **P**artially	**N**
Features: **C**ontent/ **S**ource/ **G**roup	C+S
Best predictors: **A**uthority and expertise/ **D**ata quality/ **P**opularity (**s**ource/**c**ontent)	D+Ps

2.1.4.5 Graph-based / Hybrid (Classification with Graph analyses) approach

Social networks such as Twitter can be represented as a graph which is composed of a set of nodes that are connected by a set of relationships that provide a rich set of data pieces about the social network. Within this graph, Twitter users are usually represented as nodes, where the connections between them (follows, replies, mentions and tweet) are called edges. Previous research models ignored these inter-entity relationships in Twitter however other researchers incorporated graph analysis to measure information credibility. Using a graph-based approach, the Truthy project [42]–[44] focused on tracking political memes in Twitter; it helped on detecting astroturfing, smear campaigns, and other misinformation in the context of U.S. political elections. Their approach was based on detecting the amount of similar tweets originating from an account. Using a retweet/mention graph analysis and AdaBoost with DecisionStump, SVM classifiers, their accuracy scores were high and achieved between 88% -95% based on 31 graph features (e.g. Number of nodes, Number of edges, Maximum (in, out)-degree, etc.)

Another study by Gupta, Zhao, & Han 2012 [41] proposed a credibility analysis approach enhanced with event graph-based optimization. They incorporated all the features that Castillo, Mendoza, & Poblete 2011 [13] used and included a few novel features. With some labelled data, they trained various classifiers: SVM, Naive Bayes, IBk, and J48. Then they proposed two additional steps: 1) BasicCA which performs PageRank-like iterations for authority propagation

on a network consisting of events, tweets and users and 2) EventOptCA which constructs another graph of events within each iteration and enhances event credibility values using the intuition that “similar events should have similar credibility scores”. As a result, their accuracy significantly improved to 86% compared to using decision tree classifier approach 72%. It should be noted that their classifier approach yielded lower accuracy 72% than Castillo et al. 2011’s which was 86%. Another interesting observation was that some of the used features were prominent with different datasets, but many were not. This is a valuable insight to study the effect of different datasets topics on the credibility classification.

Ravikumar et al. 2012 [45] also modelled Twitter as a three-layer graph consisting of: users (using following-follower relationships), tweets (using similarity relationships) and web pages (using PageRank). As a first step towards a complete ranking, this study only covered the tweets graph. They represented the tweets as weighted graph with tweets as vertices and edges as similarity. Content agreement between tweets was examined using Soft-TFIDF with Jaro-Winkler similarity. They used Twitter's trending topics spanning across current news, sports and celebrity gossips as their dataset and manually labelled the tweets with a relevance value and as trustworthy or untrustworthy. For evaluation, they compared relevance, trust, and time of the proposed ranking method against popular ranking of TF-IDF based on query similarity. Initial evaluations showed improvement of precision and trustworthiness by the proposed ranking and acceptable computation timings.

Yin et al. 2012 [46] proposed an adjacent users trust measurement model AUTrust applied on Weibo - China's leading micro-blogging service. They built a trust social network by determining the trust values of relationship between users computed with AUTrust. They classified the factors affecting trust between users into three dimensions: 1) similarity between users, 2) familiarity between users, and 3) users’ social reputation. They quantified trust with these dimensions. In similarity, they checked interest similarity and attributes similarity (such as age, gender, educational background and so on). While to calculate familiarity, they used both network structure to check common neighbours and common communities and interaction (such as mention, comment, retweet, email, timely chat, homepage visit, and so on). In measuring the user reputations, they referred to three factors (number of followers, following/follower ratio and quality of tweet which was the amount of user’s tweet retweeted and commented). After applying the trust model between users on 10-million-level Tencent Weibo

dataset, they concluded that trust generated by familiarity on users' interactions and users' social reputation can be used to reflect the asymmetry of trust.

Table 2-8 presents observations of this model along with the previous criteria. In addition, we will add some information about their graph structures. It should be noted that graph network metric such as number of nodes and edges, average authority of users, and density of graph may adopted by studies in this subsection. Therefore, we will add the network graph features as well in the features criterion.

Table 2-8 Evaluation of the graph-based / hybrid models

Criteria/Model	**Truthy project [42], [44], [50]**	**Gupta, Zhao, & Han 2012 [41]**	**Ravikumar et al. 2012 [45]**	**Yin et al. 2012 [46]**
Topic genera: Mixed/ Specific	S- Politics	M	M	__
Language: English/ Spanish/ Arabic/ Chinese	**E**	**E**	**E**	**C**
Level: Message/ Event/ Source	E	E	M	S
Task: Believability classification / Detecting false rumours	D	B	B	__
Labellers: Crowdsourcing site/ Authors/ Experts/ Volunteers	A+V	A	A	__
Labeller data considered: Yes/ No/ Partially	**N**	**N**	**N**	**__**
Features: Content/ Source/ Group/ Network	N	C+S+G	C	C+S
Best predictors: Authority and expertise/ Data quality/ Popularity (source/content)	Ps,c	A+D+Ps	D	A+Ps

Based on our analysis study of different credibility assessment models to purposely examine how the selected studies integrate credibility components, we argue that (labeller/user and context) components need more research attention. In terms of labeller component, we believe that labellers' data and reliability should be investigated and considered. With respect to context, we suggest to that exploring credibility assessment in other cultures and languages such as Arabic would be an enlightening research focus.

2.1.5 Twitter credibility surveys

In this section, we reviewed the different ways surveys were conducted in previous studies to identify influential credibility factors. Table 2-9 summarizes number of surveys have been carried out in this research.

Table 2-9 Twitter credibility user surveys

Kang, ODonovan, & Hollerer 2012 [15]	**Measurement:** varying users' information such as number of followers, and retweets on perceived credibility rating. **Results:** A relative rating shift between the extreme contexts.

Canini, Suh, & Pirolli 2011 [40]	**Measurement:** the effect of user name and icon, domain of expertise (on-topic, cross-topic, or off-topic), social status (followers, followings, tweets, and list memberships), and visualization such as word cloud on both explicit and implicit judgments of credibility. **Results:** The expertise factor had a strong influence on credibility judgments, and social status had a smaller influence. Additionally, visualization factor had the smallest influence on credibility judgments. Neither tweets alone nor word clouds alone provide sufficient information for participants to grant a high credibility rating to a Twitter user, but the combination of presenting specific tweets along with a summary word cloud leads to higher judged credibility.
Pal & Counts 2011 [51], [52]	**Measurement:** the effect of gender, author's name value (anonymously, non-anonymously) and type (individual, organization, topical) on the perception of quality of Twitter authors. **Results:** Ratings of authors and their content were affected either positively or negatively by author's name. Also that user names of male, organizations, and topically related to the tweet received higher ratings than those which were not.
Morris et al. 2012 [53]	**Measurement:** 1. Credibility concerns (encountering tweets, tweets topic type, tweets' features) 2. Credibility perception factors (truth, message topic, user name, user image, and reader experience and demographics (age, gender, or Twitter experience) on credibility ratings. **Results:** Participants do trust followers' tweets more than encountering tweets information by other methods. Regarding the topic type, respondents were more concerned about credibility related to news, politics, emergencies, and consumers (reviews/ offers). Features that perceived the most impact and attention from users include: followers, retweets, mentions, expertise, verification account, referencing, and similarity. They also showed that users are poor judges of the true truth value of the messages based only on content and are often biased by other information like username type and retweet. In addition, users' experience with Twitter and their demographics did not impact their ability to distinguish true tweets from false. However, experienced users rated tweets credibility higher than others as it indicated they believed Twitter to be a credible information source.
Yang et al. 2013 [54]	**Measurement:** different communities (U.S. and China) on evaluating microblog credibility. **Features Considered:** Gender, name style, profile image, location, network overlap, and message topic. **Results:** There are key differences between the two countries; Chinese users show relatively high trust and dependence on microblogs as source of information source, greater acceptance of anonymously and pseudonymously authored content, and tend to be more depend on integrate multiple metadata when evaluating microblog credibility. This implies that users' credibility perception might be cultural-dependent.
Westerman, Spence, & Van Der Heide 2012 [55]	**Measurement:** the effect of followers' number and ratio between followers and followings on ratings of trustworthiness. **Results:** They found that too few or too many followers actually make a Twitter user seem less credible. In addition, the ratio between the number of followers and the number of a user follows has an effect on the degree to a user is judged to be competent in a specific subject. That is, if one has many followers, but does not follow many others, that person is regarded as less of an expert.

As outlined above, there is variation in how credibility surveys are conducted considering the number of features presented to the participants and the monitored implications. Most of the previous surveys mainly manipulate data (i.e., user images, user names, etc.) within their experiment to measure its impact on users' credibility judgments. Therefore, it is suggested to conduct credibility survey that presents all tweet messages along with their cues/features to participants to rate the credibility of messages, and identify the features' importance. The survey results then could be used to identify the features that have more influence on credibility perception in addition to results from both the statistical and machine learning approaches.

2.2 Conclusions from Review of Literature

In this chapter, we evaluated different automatic credibility assessment methods related to UGC microblogging platform Twitter. Based on our survey, we outline the following several aspects that require further investigation:

- Although there is a variety of literature on credibility evaluation for Twitter UGC domain, most existing credibility models were based on English content, labelled by mainly western users. Indeed, credibility perception is subjective [10], [11]; and should be considered in the context of a specific community. Moreover, cultural diversity may affect peoples' attitude and preference, and how they interpret UGC information. Within this framework, exploring credibility assessment in other culture and language such as Arabic would be an interesting area of research. Most research in measuring information credibility automatically has been addressed in English language. Unfortunately, there is only one study, as far we know, about credibility measurements of Arabic Twitter content and it did not cover most of Twitter features [8], [9]. Therefore, there is a need to apply previous assessment methods and investigate their usefulness with Arabic content.
- Most of the proposed methods for credibility evaluation are based on user credibility judgment ratings. Ground truth credibility values are typically gathered using human participants, who act as information consumers and evaluators. We believe that there are differences between evaluators in how credibility perception is manifested. Attributes such as demographic profile, UGC platform experience, topic familiarity and expertise, propensity to trust, and attitude toward the topic might alter information evaluators' perceptions and can contribute to a deeper understanding about how and why users carry out their credibility judgments. Most surveyed research did not pay sufficient attention to this issue. Another

vital question that warrants inquiry (and was missing by previous studies) was evaluating crowd labellers' reliability to justify the quality of their credibility ratings as it is important to infer the true credibility labels of the tweet messages hence building more robust classification system.

- Even though previous research already proposed a feature-driven approach to assess credibility, it did not investigate the usefulness of these features in informing credibility judgments in Arabic setting. Further, there was a gap in previous research incorporating user surveys' results in identifying the credibility features importance. Our recommendation is to integrate the user survey results with the classification techniques results to help identify and correlate the credibility factors that influence credibility perception.

3 Data and Method

With reference to the first research objective, evaluating Arabic content in Twitter, a corpus of Arabic microblogging messages was required with its labelled credibility ratings in order to build the credibility model. Since no Arabic dataset existed, we confronted this problem by building a novel human annotated Arabic Twitter corpus that could be used for further research. This chapter identifies the steps needed for building the dataset and acquiring the credibility ground truth labels. The collection process and the characteristics of the newly created dataset are elaborated in the first sections of this chapter. This is followed by basic analysis of submitted credibility rating values and the collected labellers' data. The last sections are devoted to the agreement calculation for the used dataset using different settings proceeded by the procedure used to construct credibility ground truth labels using majority voting method.

3.1 Data Collection and Survey Study Design

In this section, we describe the employed methodology and the characteristics of the analysed dataset. The study was conducted using an online survey website to capture data that took place from Oct/13/2014 to Dec/10/2014 with a sample of 52 voluntary labellers. Participants were approached using invitation emails/ reminder emails with online surveys sent to different mailing lists (colleagues, friends, Saudi Universities, Saudi students' clubs, etc.) and asking them to email it to others in their mailing lists. In addition, we used social networks to distribute the user study. Participants independently evaluated and submitted 4173 credibility evaluation values for a sample of 199 unique tweet messages - 6249 tweet messages including retweets (i.e., reposted tweets by other users) - from 9 news topic categories: hard news topics such as crises, politics, health, and soft news topics such as entertainment and sports as shown in Table 3-1. A range from (5 to 37) participants evaluated each tweet message in our dataset; averages of 20 credibility evaluation values per tweet. Detailed snapshots figures of the study survey are listed in the Appendix A (Figure A-1, Figure A-2, Figure A-3, Figure A-4, and Figure A-5).

Labellers not only label the tweet messages but also annotate an assessment for the credibility features and identify cues or phrases which trigger the sense of uncertainty in the tweet messages. Online user study with the sections below is used for this step:

- **Labeller section (pre-labelling):** Participants responded to user survey questions about their demographics, Twitter usage, and familiarity different topics.
- **Labelling section:** It is the main part of the user study where participants assign credibility degree to several tweet messages covering different topics based on the provided information (tweet content and author).
- **Credibility indicators section (post-labelling):** Participants answer user survey questions to indicate the importance of different features on assessing the information credibility.

Labelling tweet messages for credibility is a time-consuming process. Typically it can take from 1 to 2 hour to answer the user survey questions, read the messages' content and assign the credibility rating scores. From a large corpus of tweet messages gathered using NodeXL Twitter Search API tool [56], random selections of tweet messages requiring labelling were placed online for volunteers' participants to be evaluated. In order to reduce the data amount and to increase the ratio of relevant and eligible data in the sample, all irrelevant, non-informative tweets are removed from the sample manually such as jokes or comments on the event.

Table 3-1 Annotated dataset

Topic	No. of unique Messages	No. of Messages (including retweets)
Topic#1: Crises - Health - Domestic - Corona virus in Saudi Arabia (فيروس كورونا في السعودية) April 2014	41	254
Topic#2: Crises - Airlines - International - Missing Malaysia Airlines flight MH370 (اختفاء الطائرة الماليزية) April 2014	20	322
Topic#3: Crises - Terrorism – International/ Domestic - Boston Marathon Explosions (تفجيرات ماراثون بوسطن) April 2013	19	217
Topic#4: Crises - Ferry transportation - International - South Korea Ferry Sinking (غرق العبارة الكورية) April 2014	18	225
Topic#5: Crises - Mine Fire - International - Mine Explosion in Turkey (انفجار منجم في تركيا) May 2014	28	191
Topic#6: Politics - Domestic - Gulf states withdraw ambassadors from Qatar (سحب سفراء دول الخليج من قطر) March 2014	23	4662
Topic#7: Health - Domestic - Diabetes in the Gulf Region (مرض السكري في منطقة الخليج) April 2014	19	283
Topic#8: Sport - International - FIFA World Cup (كأس العالم لكرة القدم) April 2014	20	81
Topic#9: Entertainment - Domestic - Arab Idol Changing Jury Panel for 2014 season (تغيير لجنة التحكيم في اراب ايدول) April 2014	11	14

In the online survey, labellers read each tweet one by one and mark the tweet credibility using 1 to 5 Likert scale where higher evaluations (5) represent higher credibility level and (1) means clearly low-credibility. The other options represent levels between these two limits; Table 3-2 below shows the credibility annotation schema.

Table 3-2 Credibility annotation schema

1	Low-credibility
2	Moderately low-credibility
3	Questionable
4	Moderately high-credibility
5	High-credibility

Experiment was conducted using two versions of annotated presentations as shown in Figure 3-1; one as “snapshotted” Twitter-presentation tweets and the other as Text-presentation tweets in order to measure the effect of appearance on perceived credibility. Indicators related to tweet content and author such as the author’s profile and number of retweets are also presented in both annotated presentations.

Figure 3-1 The two versions of annotated presentations (Text and Twitter-presentation)

All individuals who have Twitter account and know how to use it were allowed to contribute in the labelling task. Crowd labellers were almost self-selected volunteers and not trained as credibility judges. Credibility judgements were submitted independently, as each labeller in the crowd had to submit their credibility scores independently of the others. Labellers had to rely on

their own judgment while deciding on the credibility value for each tweet message and without considering any additional information; for example no training or supervision was offered.

Participants were not trained but were provided some explanations; followed by a labeller survey section where participants answered basic questions about their demographics, twitter usage and personality traits from International Personality Item Pool[10]. Then, they replied to questions about topic familiarity and interest and assigned credibility levels to several tweets covering different topics based on the provided information (tweet content and author). Lastly, they responded to questions about the importance of different features on assessing the information credibility of Twitter messages. In addition, we included an optional task concerning: believing rumours, where participants read a group of tweets and then indicated the credibility level for the discussed topic (not the tweets). The rumour used in this phase is: (منع استخدام ٥٠ اسما للمواليد في السعودية - Preventing the use of 50 names for new-borns in Saudi Arabia).

3.2 Labelling Mechanism

Building a credibility assessment model typically involves human judges referred to as "labellers" who review the tweet messages content and features and assign it to a certain class. In order to assess the credibility of tweet messages in Arabic language, we employ the idea of crowdsourcing where labellers with varied levels of expertise can explicitly express their opinions about the credibility of a set of tweets. Ultimately, "crowdsourcing is based on a simple, but powerful, concept: virtually everyone has a potential to plug in valuable information" [57]. Crowdsourcing has been used for perception-based labelling such as sentiment and relevance judgments [58] where multiple independent labels are collected from the crowd (volunteers or part-time workers) to achieve the needed labelling task. Despite the fact that labellers have varied levels of expertise, having multiple annotations would balance out the trade-off regarding quality from expert labellers. Surowiecki 2004 [59] claimed in his book that by merging opinions of individuals group, we can come up with right answer estimation better than made by any single member of the group. Another previous study by Snow et al. 2008 [60] on crowdsourcing attested that having multiple annotations from non-expert labellers is enough to reach expert quality.

[10] http://ipip.ori.org/

3.2.1 Quality of crowd labelling

Most of the reviewed studies developed and tested their models where labellers are involved to provide their individual judgments to be integrated into classification process. Due to lack of expertise information of labellers coupled with the possibility of hurried and careless labelling, the quality of annotations is always dubious. Hence, the overall quality of the labelling task depends on the reliability of the labellers. In addition, annotation process is highly subjective due to labellers' bias and cognitive heuristics such as topic knowledge and the selection of cues in making a credibility judgment [10], [11]. Therefore, to improve the quality of labelling, attempted to 1) collect larger number of credibility judgements for the same tweet message; compared to previous studies as shown in Table 3-3; in order to have more representative labelling. This redundancy allows us to implement our proposed model to conclude the correct credibility labels of the tweet messages. 2) collect and derive as much information from crowd labellers and their rating judgments as possible before it was combined to estimate the correct labels, and 3) evaluate labellers' contributions by maximizing weights of labellers with high quality labelling and diminishing labellers' weights with low quality labelling.

Table 3-3 Annotations in previous studies

Used Study	No. of Labellers	No. of Messages assigned	Final labels
Bhattacharya et al. 2012 [12]	3	2105 tweets.	Majority voting (2 out of 3)
Castillo et al. 2011 [13]	7	747 topics, each with 10 tweets.	Majority voting (5 out of 7)
Gupta et al. 2012 [14]	3	14 topics, each with 500 tweets.	Majority voting (2 out of 3)
Kang et al. 2012 [15]	145 for the user survey but no mention if their ratings used for credibility classification	591 tweets.	__ Is same tweet tested by different labellers?
Qazvinian et al. 2011 [16]	2	5 topics; 10,000 tweets (first large-scale publicly available rumour dataset).	__
Xia et al. 2012 [17]	5	289 tweet messages	__
O'Donovan et al. 2012[18]	236	1 topic; no mention for number of tweets; 6369 individual credibility assessments.	__ Is same tweet tested by different labellers?
Al-Eidan et. al 2010 [9] and Al-Khalifa & Al-Eidan 2011 [8]	3	2 topics; 29 tweets	__

In collecting credibility assessments, we also relied on experts-labellers to identify credibility levels. Three professional health experts with varying experience took part also in the credibility evaluation. These three experts will be abbreviated with the first letter of their names:

1) Expert1_A: Consultant Doctor - Department of Surgical Specialities, King Fahad Medical City.
2) Expert2_N: Consultant Doctor - King Abdulaziz University Hospital
3) Expert3_E: Demonstrator in Oral Biology Division, King AbdulAziz University, King's College London.

This means that, in addition to crowd assessments, we would depend on other judges' experts' assessment. We believe that in first step it is more related about believing the message while with the judges' assessments, it is asserting the message credibility. It also helped us validate human crowd ratings' credibility values by comparing their rating scores to ratings from reliable expert labellers. For the purpose of this study we considered the credibility assessments assigned by the experts as "correct".

3.3 Basic Analysis of Credibility Rating Values

Analyses of the submitted credibility rating values are presented in figures below. Figure 3-2 and Figure 3-3 illustrate the general credibility rating levels for all tweets' topics while Figure 3-4 shows their distribution among different topics. It is clear that most evaluation values 46.28% were within low-credibility classes {1, 2} which might indicate that Arab users are sceptical and have relatively low trust in Twitter information. As shown in Figure 3-4 there is no obvious difference on the impact of type of information to be evaluated with perception of credibility; most topics gain low credibility scores with a high percentage for the first two evaluated topics that related to crises (Corona virus in Saudi Arabia and Missing Malaysia Airlines flight MH370). A detailed credibility rating distributions for all topics is listed in Table 3-4.

Table 3-4 A detailed credibility rating distributions for all topics

	All		Topic#1		Topic#2		Topic#3		Topic#4	
1	1203	28.83%	499	35.02%	143	33.73%	77	22.58%	60	22.81%
2	728	17.45%	289	20.28%	100	23.58%	56	16.42%	45	17.11%
3	1181	28.30%	330	23.16%	102	24.06%	110	32.26%	86	32.70%
4	797	19.10%	207	14.53%	68	16.04%	68	19.94%	59	22.43%
5	264	6.33%	100	7.02%	11	2.59%	30	8.80%	13	4.94%
	Topic#5		**Topic#6**		**Topic#7**		**Topic#8**		**Topic#9**	
1	109	24.55%	107	28.16%	74	20.67%	80	21.92%	54	31.21%
2	47	10.59%	72	18.95%	65	18.16%	30	8.22%	24	13.87%
3	126	28.38%	125	32.89%	100	27.93%	141	38.63%	61	35.26%
4	117	26.35%	56	14.74%	95	26.54%	101	27.67%	26	15.03%
5	45	10.14%	20	5.26%	24	6.70%	13	3.56%	8	4.62%

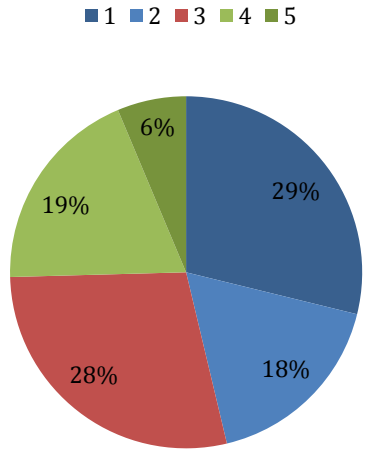


Figure 3-2 Credibility rating values - 5 classes

To gain a clear differences between low-credible messages and high-credible messages, and to ease the classification process later, the five credibility classes grouped into three classes [18], [61]. Rating {1, 2} values which are more related to low-credibility messages is combined in same class called {1}; all {3} values in questionable class are changed to {2}; and finally {4, 5} classes for high-credibility messages clustered in same class called {3}. Credibility rating values distribution was also obtained with 3-class rating schema applied on the same data as shown in Figure 3-3.

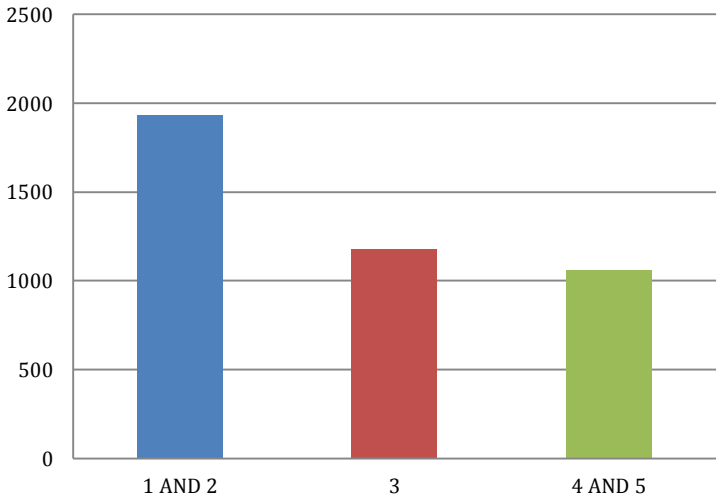


Figure 3-3 Credibility rating values - 3 classes

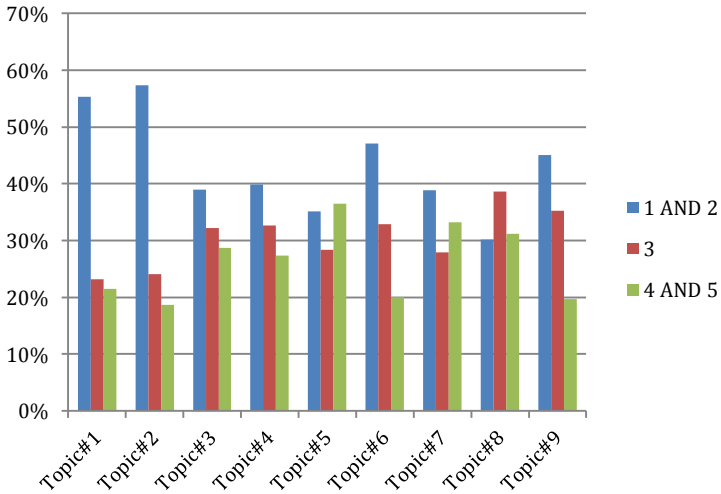


Figure 3-4 Credibility rating values by topic - 3 classes

Different annotation presentations have been used to measure the effect of messages' appearance on perceived credibility. Figure 3-5 shows the credibility ratings distribution among the two different presentations: "snapshotted" Twitter-presentation and Text-presentation. As shown from the graph, the messages' appearance produced a noticeable impact on credibility perception. Most of the low-credibility ratings {1, 2} were within the "snapshotted" Twitter-presentation while other Text-presentation has reasonable balanced scores from all credibility classes.

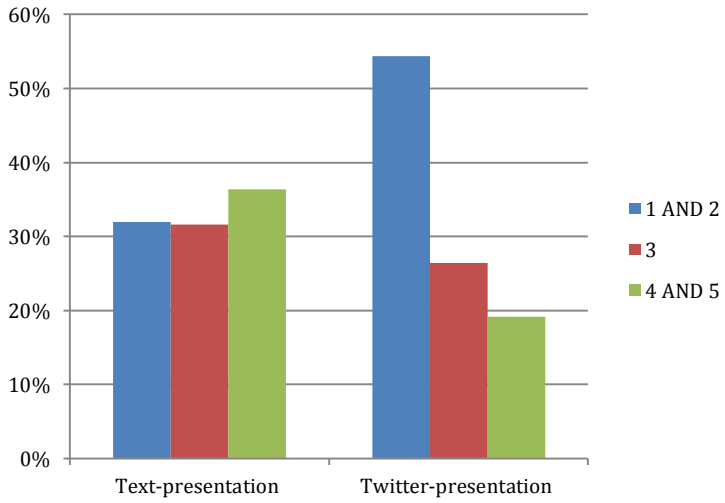


Figure 3-5 Credibility rating values by presentation - 3 classes

A total of 52 labeller participants contributed to the tweet messages credibility evaluation; they engaged in a total of 4173 credibility assessments. The labellers were free to judge as many tweet messages as feasible, consequently not all messages yielded the same number of assessments. The participants taking part in this study were 32 females and 16 males with an average age range between 22-44 years old. Participants were mainly from Saudi Arabia and holders of a superior level of education - 58% have a postgraduate degree and 25% are Bachelor graduates. They originated from a variety of educational and occupational backgrounds; medical, speech pathology and audiology, computer science and information system, languages, accounting, economic, marketing, financial, public policy, media, communications, and engineering. It is worth mentioning that three participants were experts in the healthcare field.

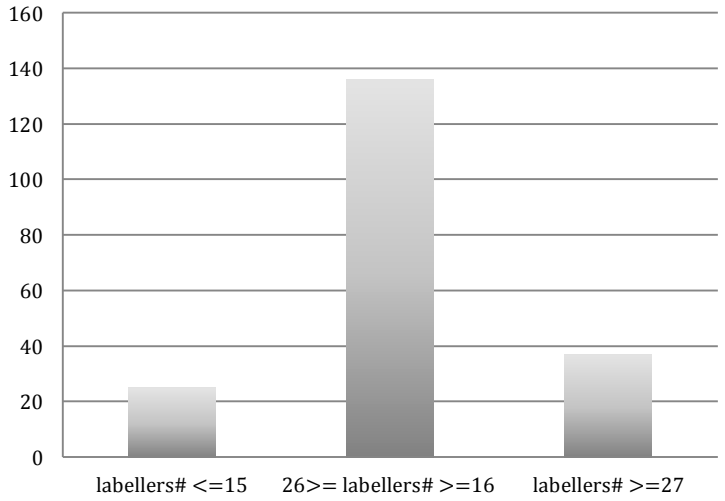


Figure 3-6 The distribution of labellers across rated tweets

Every participant could annotate more than one topic's tweets; a range from (5 to 37) participants evaluated each tweet message in our dataset; averages of 20 credibility evaluation values per tweet. Figure 3-6 shows the distribution of labellers across rated tweet messages where Figure 3-7 resembles the distribution of rating count average across topics which approximately following a power-law distribution. It means labellers were initially eager to rate the tweets for first topics (topic#1) but most of respondents did not complete the remaining topics, so the number of labellers diminishes as the number of topics increase.

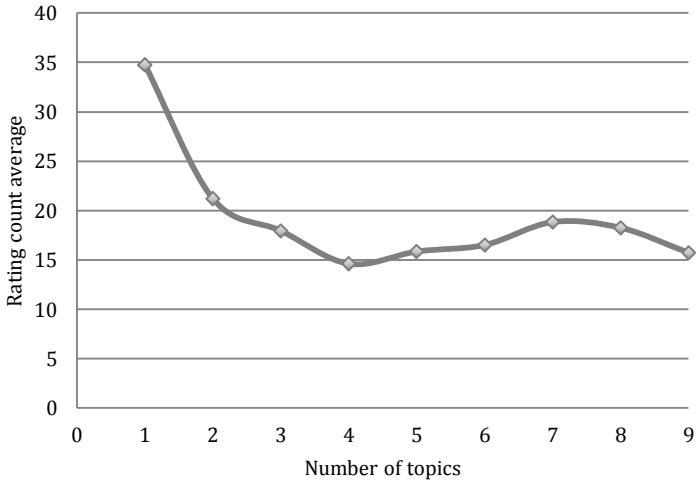


Figure 3-7 The distribution of rating count average across topics

3.4 Basic Analysis of Labellers' Data

In order to satisfy one of the thesis objectives, which is to extract as much information about labellers' as possible, the first section on the online user study was designed to gather data about labellers' demographics, Twitter usage, and different topics familiarity. A number of statistical graphs are exhibited below to examine labellers' traits and its impact on credibility perceptions. We statically study some of the labellers' traits proposed by Kakol et al. 2013 [34], Flanagin and Metzger 2003 [62] and Fogg et al. 2001 [26], [35] for Web sites credibility in addition to extra proposed traits related to topic familiarity and interest.

3.4.1 Labellers' age

Labellers were divided into four age categories and the majority were within [22-34] and [35-44] age categories. As shown in Figure 3-8, the distribution of the credibility ratings across different age groups revealed that Arab youth under 35 years old have relatively higher trust in Twitter information. This finding did not agree with results from Fogg et al. 2001 [26] who

claimed that the younger generation tended to be more critical and stringent on their credibility judgments.

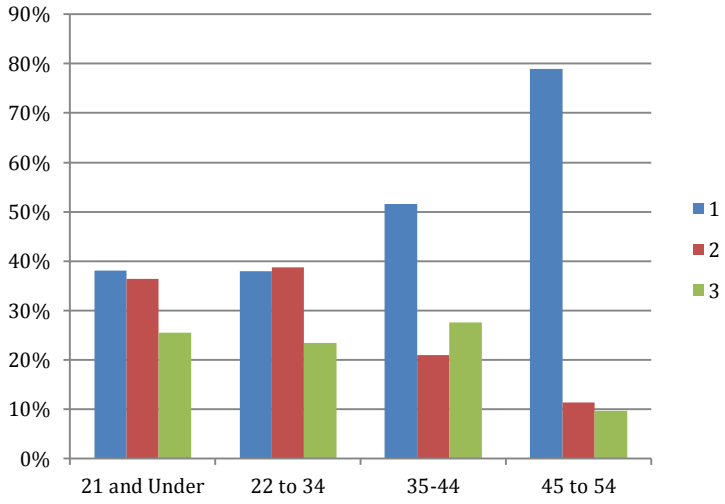


Figure 3-8 The distribution of ratings across labellers' age categories

3.4.2 Labellers' gender

The majority of the labellers in the study were female (66%), as previously stated. Figure 3-9 suggested that although both genders generally produced more low-credibility ratings scores, there was a considerable difference in the distribution of the credibility ratings between the two genders. Males were harsher in their credibility judgments as they offered less high-credibility scores to tweet messages compared to other credibility classes. This finding correlates with the results by Fogg et al. 2001 [26] who also reported that men assigned more lower credibility ratings compared to women.

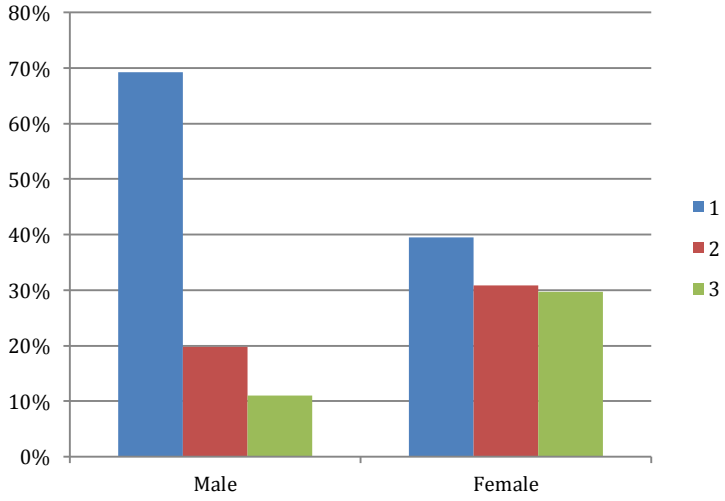


Figure 3-9 The distribution of ratings across labellers' gender categories

3.4.3 Labellers' education

More than half of the labeller participants possess a superior level of education – in fact, 58% hold a postgraduate degree and 25% are Bachelor graduates. It is noticeable from Figure 3-10 that labellers with higher levels of education, also tend to assign more low-credibility scores and are more confident and decisive in their judgments (less "questionable" credibility ratings). On the whole, labellers from different educational backgrounds perceived credibility almost the same way which is also reported by Fogg et al. 2001 [26] who did not find a significant difference in credibility perceptions between participants based on their education level measure.

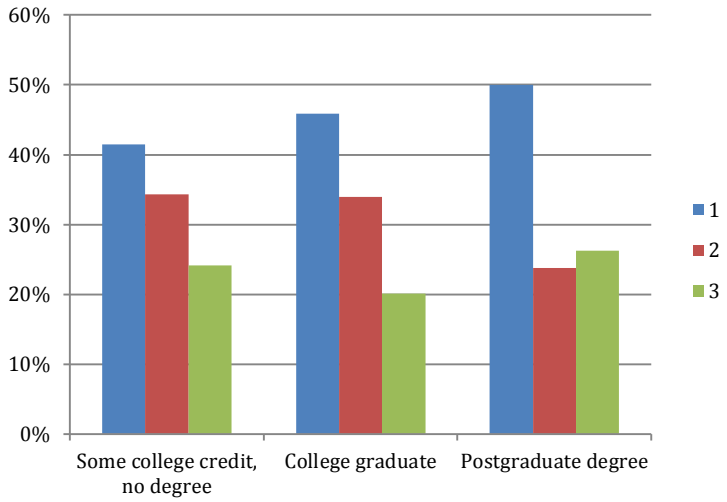


Figure 3-10 The distribution of ratings across labellers' education categories

3.4.4 Labellers' Twitter features and usage

In regard to labellers' experience with Twitter, most labellers (more than 80%) had created their twitter accounts at least 2 years ago at the time of user study. More than 60% of them check news updates on Twitter at least once per day. However the majority (80%) only tweet or retweet messages on Twitter at most several times per week (40% of them at most once per month). With respect to their Twitter features and influence, 67% of labellers have followers less or equal to 200 and all of them follow less than 1000 Twitter accounts (more than 60% follow at most 200 accounts). Labellers' Twitter features were mainly assessed on a five-point and four-point scales depending on number of items presented in the user survey. Based on Labellers' answers about their Twitter features and the frequency of Twitter usage, they were divided into three categories (Low, Average, and High). The majority of labellers were classified as average Twitter users. Credibility ratings' distribution shown in Figure 3-11 showed that labellers who

tend to use Twitter more often and have more influential Twitter features tend to assign high-credibility ratings.

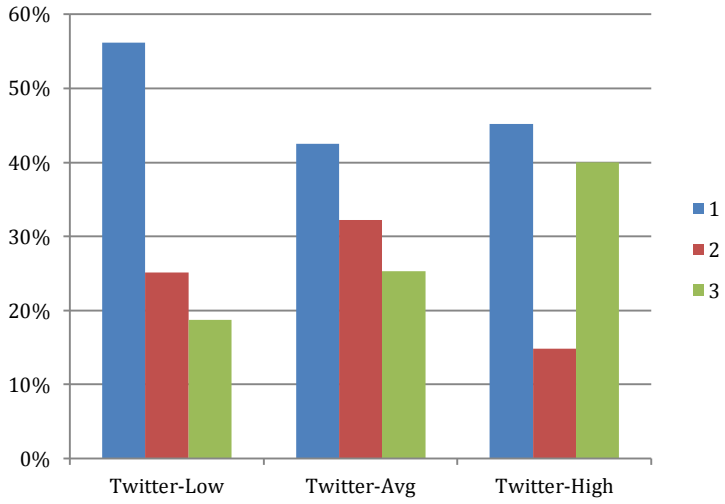


Figure 3-11 The distribution of ratings across labellers' Twitter features and usage

3.4.5 Labellers' personality trust trait

As part of our data collection of some personality labellers' features, we measured labellers' trust characteristics using the trust scale from the International Personality Item Pool [63]. IPIP is a public-domain personality catalogue containing items developed among scientists worldwide to measure Individuals' personalities and differences. Labellers' trust feature was assessed with five items on a five-point scale start from "Very Inaccurate" to "Very Accurate". Items included in the study were: 1) "Believe that people are basically moral", 2) "Suspect hidden motives in others", 3) "Believe that people seldom tell you the whole truth", 4) "Act comfortably with others", and 5) "I'm wary of others". Items were averaged to form an overall trust score. Most labellers 45% who provided answers to this part of the study were classified as having average willingness to trust. Figure 3-12 shows that there is no significant difference in the credibility ratings distribution between the different categories although labellers with high tendency to trust assigned to some extent more high-credibility evaluations compared to other labellers.

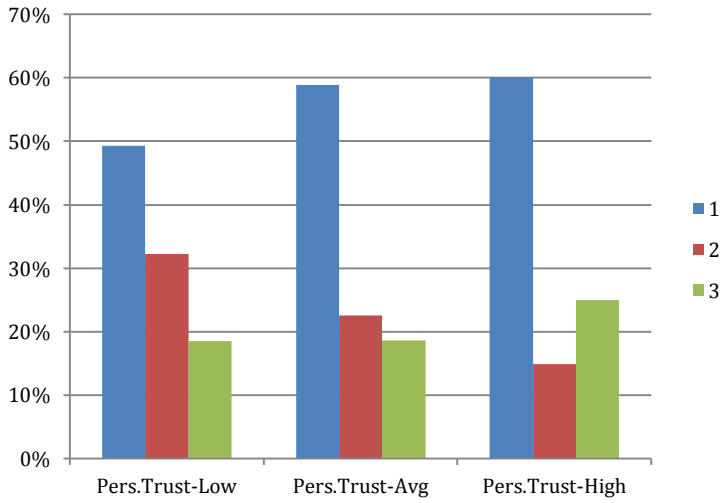


Figure 3-12 The distribution of ratings across labellers' personality trust trait

3.4.6 Labellers' topic familiarity and interest

Questions regarding topic familiarity, topic interest and attitude towards source or topic were included in the user survey. Since responses in the questionnaires may be subjective, such as assessing one's familiarity with the topic, we included control questions to increase the likelihood for objectivity. For example, in crises-topic, in addition to asking "How familiar are you with the "Missing Malaysia Airlines flight MH370" news event?" as an example for topic familiarity, we also included a control question "The Missing Malaysia Airlines Flight MH370 was flying from Kuala Lumpur to: 1) Singapore, 2) Beijing, 3) Hong Kong, 4) Sydney". Results indicated that the majority of labellers have low to average topic familiarity. Figure 3-13 implied that participants with higher topic familiarity and interest assigned a notable low-credibility rating scores compared to other groups.

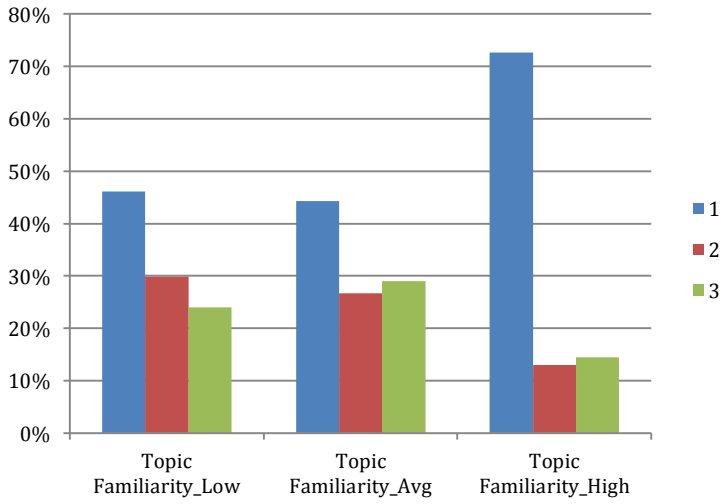


Figure 3-13 The distribution of ratings across labellers' topic familiarity and interest

.

3.5 Agreement Calculation and Interpretation

For a manually annotated dataset that consisted of multiple independent credibility judgments for the same tweet messages, it was prudent to measure the extent to which labellers agree when rating the same set of messages. Moreover, this technique can be treated as a sort of a statistic measure to estimate data reliability for model reproducibility. There are a number of statistics that can be used to measure agreement among labellers: Percent agreement; Scott's pi (π) 1955 [64], Cohen's kappa (κ) 1960 [65], Fleiss' kappa (κ) 1971 [66], and Krippendorff's alpha (α) 2004, 2012 [67], [68]. Although Cohen's kappa (κ) is the standard popular measurement, in this research, we applied the Krippendorff 's alpha (α) measurement to assess the inter-labellers agreement. Krippendorff's alpha (α) has an advantage among other measurements due to its widespread ability to be used for any number of labellers, different kinds of data, including incomplete or missing data [67]. In addition it has been used in domains like opinion retrieval [69] and computational linguistics [70] where subjectivity in judgments is applied.

3.5.1 Krippendorff's alpha (α)

Due to its ability to calculate agreements when missing data are present, which is relevant to our dataset, Krippendorff's (α) [67] was utilized in this study to measure the inter-labellers agreement. Unlike kappa (κ), Krippendorff's alpha (α) does not consider observed and expected agreements but considers observed and expected disagreements. α = 1.0, represents exact agreement, α = 0.0 represents exact disagreement. Krippendorff's alpha (α) can also return negative values, meaning that agreements are below chance [71], [72].

$$\alpha = \frac{D_e - D_o}{D_e} = 1 - \frac{D_o}{D_e}$$

D_o is the observed disagreement between labellers (among credibility evaluation values) and D_e is an estimation of the possible chance disagreement. For detailed computational steps, Krippendorff [73] provided 4 different examples which cover different data options 1) binary data, two labellers, no missing data, 2) nominal data, two labellers, no missing data, 3) nominal data, any number of labellers, missing data. 4) All metrics, any number of labellers, missing data.

Below is an example of the first option which considered the simplest form: binary data for 10 observations, two labellers, with no missing data.

Observation#1: *1* *1* *disagreement*
Observation#2: *1* *1*
Observation#3: *1* *0* *disagreement*
Observation#4: *0* *0*
Observation#5: *0* *0*
Observation#6: *1* *0* *disagreement*
Observation#7: *0* *0*
Observation#8: *0* *0*
Observation#9: *0* *1* *disagreement*
Observation#10: *0* *0*

	0	***1***	∑
0	*10*	*4*	*14*
1	*4*	*2*	*6*
∑	*14*	*6*	*20*

Total disagreements (decision pairs) = 4
Total coded values of 1 = 6
Total coded values of 0 = 14
Total coded values = 20

General form of Krippendorff's alpha for a binary variable:

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - (n-1)\frac{O_{01}}{n_0 n_1}$$

Worked example of that equation for this example:

$$\alpha = 1 - (20-1)\frac{4}{(14)(6)} = 0.095$$

Computations of Krippendorff's (α) are performed using SPSS macros written by Andrew Hayes[11]. Table 3-5 reports the computed inter-labellers agreement values using Krippendorff's alpha (α) for 5 and 3 credibility classes taking in consideration different settings:

- **Different labellers:** To examine inter-labellers agreement values for crowds and experts.
- **Different presentations:** To investigate inter-labellers agreement values for labelling with Text-presentation and Twitter-presentation.
- **Different topic types:** To compare agreement values obtained with different topics, ratio of messages per topic, and ratio of labellers per topic.

[11] http://www.afhayes.com

Table 3-5 Krippendorff's alpha (α) values for different settings

Krippendorff's alpha (α) for different settings	No. of Messages	No. of Labellers	5-Classes	3-Classes
Different labellers, different topic types, and different presentations	199	39	0.1176	0.1180
Crowds, different topic types, and different presentations	199	37	0.1128	0.1150
3 Health experts, Topic#1+Topic#7, and Twitter-presentation	57	3	0.1203	-0.0034
Different labellers, different topic types, and Twitter-presentation	174	30	0.0853	0.0611
Crowds, different topic types, and Twitter-presentation	174	28	0.0752	0.0801
Crowds, different topic types, and Text-presentation	199	9	**0.2142**	0.1981
Crowds, Topic#1+Topic#7, and different presentation	60	37	0.1285	0.1278

Table 3-6 reports the inter-labellers agreement values using Krippendorff's alpha (α) for different topics using 39 different labellers, and different presentations.

Table 3-6 Krippendorff's alpha (α) values for different topics

Krippendorff's alpha (α) for different topics	No. of Messages	Topic Type	3-Classes
Different 39 labellers with different presentations	41	**Topic#1:** Crises - Health - Domestic	0.1229
	20	**Topic#2:** Crises - Airlines -International	0.0648
	19	**Topic#3:** Crises - Terrorism - International/ Domestic	0.0965
	18	**Topic#4:** Crises - Ferry transportation - International	0.0207
	28	**Topic#5:** Crises - Mine Fire - International	0.0472
	23	**Topic#6:** Politics - Domestic	0.1573
	19	**Topic#7:** Health - Domestic	0.0758
	20	**Topic#8:** Sport - International	-0.0040
	11	**Topic#9:** Entertainment - Domestic	0.1107

Several general observations can be drawn from the preceding tables:

- Generally, the level of agreement among the labellers was ranged from 0.1176 – 0.1180 for 5 and 3-credibility classes; which insinuated that participants have a "slight" inter-labellers agreement based on interpretation by Landis & Koch 1977 [74]. They suggested subjective guidelines for interpreting kappa-like measures (which includes Krippendorff's alpha α) as followed: $\kappa \leq 0$ indicates poor agreement, $0 \leq \kappa \leq 0.2$ indicate slight agreement, $0.2 \leq \kappa < 0.4$ indicate fair agreement, $0.4 \leq \kappa < 0.6$ indicate moderate agreement, $0.6 \leq \kappa < 0.8$ indicate

substantial agreement, $0.8 \leq \kappa \leq 1$ indicate (almost) perfect agreement. However, Krippendorff 2004 [60] contends more conservative interpretations suggesting that a κ or α value of .80 as a threshold for firm conclusions, and a value of at least 0.67 is sufficient for drawing tentative conclusions. In this research, we will rely on Landis & Koch 1977 [74] interpretations and we assert it is reasonable to apply these limits as it is known that alpha values α are usually smaller than kappa κ.

- It is observed from Table 3-5 that the alpha (α) values are more consistent and less affected by changing the number of credibility classes or labellers which makes it a reliable measurement. Also, alpha (α) computed values confirm that the best level of agreement 0.2142 was obtained by crowd labellers using text annotated presentation.
- An astounding outcome involves the poor agreement value obtained by experts compared to the crowd labellers. An alpha α value of $-0.0034 \leq 0$ which indicates poor agreement is obtained by experts where a value of $0 \leq 0.1278 \leq 0.2$ which indicate slight agreement is obtained by crowd. One possible explanation is that individuals (including experts) tend to have a strong bias, and combining more multiple diversity labels within may reduce the effect of labeller bias and, hence improve the quality of the data.
- In regards to topics, Table 3-6 displayed that the best inter-labellers agreement values were achieved by topics that cover domestic issues. Participants agreed with assigning similar credibility scores to local domestics which they maybe more familiar with.

3.5.1.1 Number of labellers

This section investigates to what extent the stability of the inter-labellers agreement values depends on the number of labellers. As far as we discern, there is no clear indication in the literature about a recommended number of labellers (for opinion labelling tasks) to reach a stable agreement value. Therefore, an experiment was conducted to investigate the influence of the number of labellers on the inter-labellers agreement values. The experiment start with computing the alpha agreement value for two randomized labellers' contributions, then for each iteration, we add an additional randomized labeller's ratings and re-compute the alpha value between all labellers; the iteration ends after computing the alpha value for all participated labellers. This experiment was repeated 5 times and results are illustrated in Figure 3-14.

As shown from Figure 3-14, we concluded that adding the contributions of at least 15 labellers is required to have stable agreement values; giving a low agreed dataset. This finding is evidence that including more annotations reduce the labellers' bias. It should be noted that Krippendorff's [70] recommended to start at least with three labellers to obtain usable labelling data. For future work, we will attempt to incorporate all combinations of labellers.

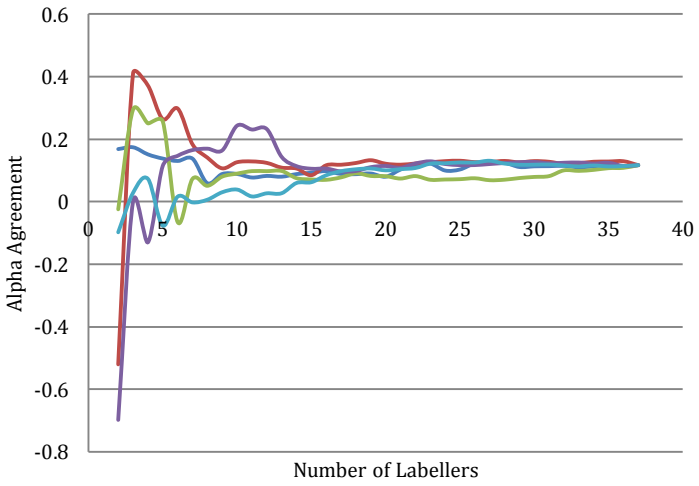


Figure 3-14 Impact of the number of labellers on the inter-labellers agreement values

3.6 Constructing Credibility Ground Truth

Based on different ways to construct the ground truth labels, Allahbakhsh et al. 2013 [75] have classified the most common techniques as follows:

- **Expert labelling:** domain experts annotate the data or at least check the contribution quality.
- **Crowd agreement:** crowds independently provide the same labels for the desired task.
- **Gold Standard:** labels are compared to approved defined labels.
- **Majority voting:** assign the label with the most votes as the correct label. It is the best applied method by credibility models research as previously shown in Table 3-3 and will also be activated in this research.
- **Labellers' evaluation:** assess contributions based on the labellers' reliability.
- **Real time support:** provide more instructions and guidelines to crowd in real-time to help them increase contribution quality.
- **Workflow management:** design workflow system to monitor the contribution quality.

3.6.1 Majority voting

In previous Twitter credibility studies, the generally accepted and common approach to estimate final labels from independent labellers is the simple majority voting [12], [13], [14]. It is based on simple easy concept: the credibility class that receives the maximum number of votes is accepted as the final aggregated label for that tweet message, i.e. the one that most labellers agree with. In this option, all labellers are considered reliable and have equal weights. A main draw with majority voting is with the cases where multiple labels receive an equal number of votes as in our dataset. Table 3-7 state the percentage values of majority rating class compared to other classes in our dataset.

Consequently, to address this drawback, we proposed labellers' weighting mechanism that aggregates all labellers' ratings with respect to their weights. Most included studies which encounter messages with conflicting credibility scores resolved by an extra judgment [12], labelling them "unsure" [13], or just discarding them [14]. However, in this research, we introduced an assumption of unequally reliable labellers where labellers are weighted to capture the reliability of each labeller. Along with this assumption all these controversial messages are considered and evaluated depending on labellers' weighs.

Table 3-7 Ratio of majority rating class compared to other classes

Ratio of majority voting class	5-Classes	3-Classes
Maximum percentage of majority voting class	75 %	100 %
Minimum percentage of majority voting class	21.62 %	33.333 %
Average percentage of majority voting class	40 %	52.655 %

The following methods formulate the credibility labels vectors using majority voting decision considering cases where multiple labels receive an equal number of votes:

- *Select the class with the maximum number of votes, but in case of having more than one, these options are considered:*
  - *Class 2 is selected - Maj_Class2 (Questionable class)*
  - *The lowest class is selected - Maj_Low;*
  - *The highest class is selected - Maj_Hi;*
  - *Check the total sum of the neighbours' cells and class with highest summation is selected - Maj_N.*

To access the more representative ground truth vector and also to identify the expert who agrees predominantly with the crowds, Krippendorff's alpha (α) agreements values between the

different ground truth vectors and the three experts participants (Expert1_A, Expert2_N, and Expert3_E) are computed as shown in Table 3-8.

Table 3-8 Agreement between majority voting ground truth vectors and experts using Krippendorff's alpha

Expert/ GT vectors	Expert1_A	Expert2_N	Expert3_E	Average
Maj_Class2	**0.6055**	0.0418	0.2100	0.2858
Maj_Low	**0.5831**	0.0814	0.1674	0.2773
Maj_Hi	**0.6075**	0.0294	0.2343	0.2904
Maj_N	**0.6055**	0.0418	0.2100	0.2858

Another simple measure to compute the agreements is the agreement percentage which is the proportion of matching values between two vectors. Even though percent agreement has garnered a lot of criticism and rejection [65] as it doesn't correct for agreements that would be expected by chance and therefore overestimate the level of agreement, many researchers continue to report the percentage agreement method in their studies [76], [77] due its ease with computation and interpretation practices. For more than two labellers, this calculation requires average pairwise percent agreement, in which the agreements of all possible pairs are calculated and then averaged. In this research, we also calculated the percentage agreements for all labellers including experts using 3 credibility classes' schemas and an average percentage agreement with value 0.43 have been reached. In addition percentage agreements between each expert and ground truth vectors was also calculated and illustrated in Table 3-9.

Table 3-9 Agreement between majority voting ground truth vectors and experts using percentage

Expert/ GT vectors	Expert1_A	Expert2_N	Expert3_E	Average
Maj_Class2	**0.6316**	0.7018	0.4737	0.6024
Maj_Low	**0.6491**	0.7368	0.4561	0.6140
Maj_Hi	**0.6667**	0.7018	0.4561	0.6082
Maj_N	**0.6316**	0.7018	0.4737	0.6024

With reference to the tables above, **Expert1_A** has the highest agreement level and best overlap from both the percentage and alpha agreements which makes him/her the more representative for crowd judgements. In addition by comparing the distribution of the credibility ratings between experts and the crowd credibility scores, **Expert1_A** showed relatively a similar distribution as the crowd compared to other experts as shown in Figure 3-15.

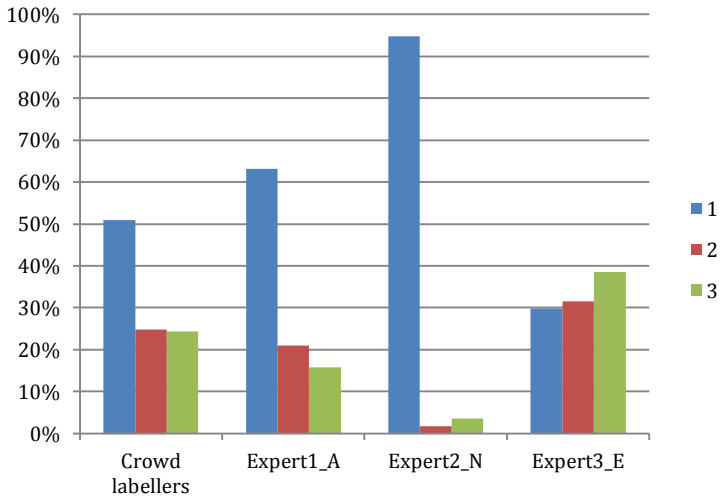


Figure 3-15 The distribution of credibility ratings across crowd and three experts

In regards to the selected ground truth vector, **Maj_Hi** has the highest agreement level with experts and best overlap from both the percentage and alpha agreements.

3.7 Conclusions from Data and Method

This chapter described the employed methodology and the characteristics of the analysed dataset along tow fundamental principles:

- First, it presented numerous statistical graphs that examined both, the collected credibility rating values and the labellers' traits. Based on these analyses and graphs, we summarized the following conclusions:
    - It is evidence that Arab users have low level of trust towards Arabic news posted via Twitter as most credibility evaluation values were documented within low-credibility classes.
    - In regards to labellers' data, there are overall differences between classified groups. However, these differences were not significant as all labellers from different traits' groups generally assigned low-credibility scores regardless of their demographic, Twitter usage, and topic familiarity. The resulting differences can be summarized as following:
        - Arab youth labellers have relatively higher trust in Twitter information.
        - Males were more suspicious towards online Twitter content; they assigned more low-credibility scores as compared to their females counterparts.

- Labellers with higher education level tend to assign more low-credibility scores and were more confident and decisive in their judgments- less credibility judgements with class {2}.
- Labellers who are more familiar with Twitter and have more influential Twitter features tend to assign higher credibility ratings.
- Labellers with a high tendency to trust trait assigned to some extent more high-credibility evaluations compared to other labellers.
- Also labellers with higher topic familiarity and interest assigned a notable number of low-credible rating scores compared to other groups.

- Later on this chapter, the agreement calculation between labellers is introduced and evaluated on the used dataset in different settings. In addition, different methods based on majority voting were formulated to construct the credibility labels vector considering cases where multiple labels receive an equal number of votes. As a result of the data from these sections, several observations and conclusions are presented below:
  - Generally, there is quite disagreement between labellers on assessing the credibility of tweets. Labellers have a “slight” agreement value; an explanation to this could be that credibility judgments are generally understood as opinion labelling which prone to high subjectivity. In addition, it is worth mentioning that crowd labellers are not monitored or even trained as credibility judges, also assessing the information credibility is a challenging task since “credibility” is a complex concept that is based on multiple dimensions. Besides crowd labellers have diverse backgrounds that might influence their credibility judgments in the annotation task.
  - Message labelling presentation has a slight influence on the credibility perception and inter-labeller agreement as presenting messages with text annotated presentation lead to reasonable balanced scores from all credibility classes and more agreement values between crowds. However, the number of rating classes has no impact as the agreement values are more consistent by changing the number of credibility classes.
  - Number of labellers has an influence on the agreement value, and to reach a stable agreement value, at least 15 labellers’ contributions were recommended to be added. It should be noted that combining more multiple diversity labels within may have the effect of reducing labeller bias and thus improving the quality of the data.

- Labelling agreement from the crowd non expert outperformed expert annotations using a variety of settings. This is not the only case were experts disagree on labelling, study by Al-Eidan et al. 2010 [9] and Al-Khalifa et al. 2011 [8] to evaluate Arabic tweets faced the situation where only 2 of the 3 experts reach to reasonable kappa (0.6) agreement. Usually in cases where labelling opinion task received multiple labels from different experts, disagreements among the experts were common because of their biases, expertise and individual differences. To sum up, experts are also individuals and subject to bias according to their different experiences.

- In this chapter, we also identified the expert who represents the crowd by computing the alpha and percentage agreements between the labelling obtained by experts and labelling using crowd majority voting. In addition by comparing the distribution of the credibility ratings between experts and the crowd credibility scores, this identified expert presented a similar distribution to the crowd compared to other experts.

4 Labellers' Evaluation and Weighting

The labelling phase requires gathering credibility rating scores for a collection of tweet messages from independent labellers with anonymous levels of expertise. It is well known that a high quality set of training data is a necessary condition for the construction of an effective prediction model. Therefore the labeller's reliability is acknowledged as an important factor affecting the prediction model. Due to the lack of expert data of crowd labellers, and the increased likelihood of carelessness behaviour of some labellers while doing the annotation task, the quality of contributions remains questionable. In addition, the crowd labellers have diverse backgrounds that might influence their credibility judgments in the annotation task. The credibility of a tweet message is subjective, based on the labellers' viewpoint, cultural background and personal experiences. Therefore, for all these reasons, we expect to have a noisy labelled dataset resulting in the probability for intensified labellers' disagreements.

According to our evaluation of current automatic credibility assessment models, these issues have been largely ignored; and while a paucity of research exists on how to characterize and handle labellers' disagreement, a second challenge looms of how to incorporate labellers' disagreement to build a better training dataset. In another words, the question is how to optimally assign an objective final credibility score to each tweet message with respect to disagreements between labellers. As the overall quality of the annotation task depends on the reliability of the labellers, this study looks at the reliability of crowd labellers in generating annotated dataset corpus and makes an assumption of labellers with varying levels of reliability. Effective solutions to inconsistent credibility ratings must be developed to produce more objective rating scores of the same tweet messages. This study propose a credibility model takes the labellers' ratings disagreements into consideration when deriving the credibility labelling in order to generate more objective rating scores. This chapter is devoted to evaluating the quality of the crowd labellers' credibility ratings using the following measures: 1) similarity and consistency, 2) accuracy, 3) agreement, 4) majority consensus, and 5) propensity to trust. These measures are utilized to maximize weights of labellers with high quality labelling and diminish labellers' weights with low quality labelling. In order to evaluate proposed weighting measures, we compared the derived labelling by proposed measures with expert's labelling. The idea was to recalibrate labellers' contributions to align with expert labelling. We determined that the proposed method is a promising idea which recognizes reliable labellers who will gain

higher weights and exhibit a superior credibility judgment that are more correlated to experts' credibility values.

4.1 Quality of Ratings' Measurements and Algorithms

In this study, we are aiming to identify comprehensive labellers' ratings evaluation framework encompassing a set of quantitative measures and iterative algorithms which provides an estimated weights of the labellers based on the quality of their credibility ratings. Therefore, we introduced the concept of labellers' reliability using multidimensional models to weight crowd labellers in order to indicate to what extent a labeller credibility ratings 1) correlate, 2) distance, 3) consensus, and 4) agree with other labellers' ratings and with the average crowd rating. The applied method uses a parallel set of measures to compute a weight for each labeller from the crowd which reflects his/her level of reliability. The basic idea is to maximize the weight of high quality labellers and reduce the influence of unreliable crowd labellers. Proposed measures are used as a metrics to identify both: the pairwise relation between labellers rating values and between the labellers' rating values and the average rating values. Our measures capture similarity, accuracy, agreement, majority consensus, and propensity to trust factors. The main strategy is to calculate the labellers' weights in two steps: in the first step, we calculate the weights using each measure; in the second step, we aggregate such estimated multiple measures in order to create the final weights ranks of evaluated labellers. Table 4-1 displays a listing of the primary measurements followed by Table 4-2 for the used notations to compute these measurements.

Table 4-1 The main used measurements

Measure	Purpose	Used techniques
Similarity and Consistency	Demonstrate if labeller's rating scores correlate with other participated labellers and average rating	cosine similarity, pearson's correlation coefficient, extended jaccard coefficient, intraclass correlation.
Accuracy	Demonstrate the distance between labeller's rating scores and other participated labellers and average rating	pairwise differences, average absolute deviation, normalized deviation, variance, variance by topic, standard deviation
Agreement	Demonstrate if labeller's rating scores agrees with other participated labellers	percent agreement, alpha agreement
Majority Consensus	Demonstrate if labeller's rating scores are closer/ consent to the ratings of the majority of the participated labellers	exact match, class ratio, normalized class ratio
Propensity to trust	Demonstrate if labeller's has low or high propensity to trust.	average absolute deviation

After computing labellers' weights using each measure, we standardize the range of weight values and normalize it between 0 and 1 using min-max normalization; the following equation describes the normalization:

$$Norm\left(W_{l_j}\right) = \frac{W_{l_j} - min(W_l)}{max(W_l) - min(W_l)}$$

Normalization within the context of this thesis refers to mapping computed labellers' weights into a new numerical range between [0, 1] to allow for comparisons of the used measurements.

Finally, derived final labels to tweet messages are estimated using the labellers' weighted scheme. Each tweet message is labelled with the credibility class that corresponds to the maximum aggregated labellers' weights taking into account the use of majority voting method to identify the ground truth value for each tweet. To validate the computed measure, we compared the derived labelling by proposed measure with expert-labelling; in particular with **Expert1_A** labelling as this participant was identified as the most representative expert (refer to chapter 3 – section 3.6.1). For the purpose of this study we considered the credibility assessment as the one compiled by the experts as "correct", or "truthful", and the assessment completed by volunteers' crowd as subjective. We used mainly Krippendorff's alpha agreement for validation in addition to other correlations measures. In the next subsections, we introduce the five evaluation metrics, and offer the applied formulas to compute them.

Table 4-2 The main notations

m	*Number of labellers*
$L[m]$	*vector of* m *labellers;* $L[m] = \{ l_1, l_2, l_3 \dots \dots . l_m \}$
n	*Number of tweet messages*
$T[n]$	*vector of* n *tweets;* $T[n] = \{ t_1, t_2, t_3 \dots \dots . t_n \}$
p	*Number of topics*
p_{l_j}	*Number of topics rated by labeller* l_j
n_{l_j}	*Number of tweet messages rated by labeller* l_j
$n_{l_j p_k}$	*Number of tweet messages rated by labeller* l_j *in topic* p_k
m_{t_i}	*Number of labellers who rated tweet message* t_i
m_{p_k}	*Number of labellers who rated topic* p_k
$CR[n.m]$	$n.m$ *integer rating matrix;* $CR[t_i, l_j]$ *denotes the Credibility Rating given to tweet message* t_i *by labeller* l_j
$AVGT[n]$	*Average rating vector;* $AVGT[t_i]$ *denotes the average of all labellers 'ratings who rated tweet message* t_i
$SDT[n]$	*Standard deviation rating vector of all labellers 'ratings who rated tweet message* t_i
$\bar{X}$	*Average value of any vector X values*
$\|X\|$	*Absolute value of any number X*

$ClassNo$	*Number of credibility classes*
$ClassC_j$	*The credibility values' count for j credibility class for tweet message* t_i
$MajClass_{t_i}$	*The class number have the maximum credibility values' count for tweet message* t_i
$Class[ClassNo]$	*vector of ClassNo contains the rating credibility classes given to tweet messages;* $Class[3] = \{1, 2, 3\}$
$ClassC[n.3]$	$n.3$ *matrix contains the credibility values' count for each credibility class and for each message tweet*
$TLabel_{t_i}$	*Estimated label for tweet message* t_i
W_{l_j}	*Weighted value of labeller* l_j

4.1.1 Similarity and consistency model

Credibility evaluations might be conducted using careless labellers who might give arbitrary credibility scores, which do not correlated with the others' assessments. The similarity and consistency model quantify the rating similarity between two vectors of rating score values. It indicates the degree tendency of two rating values vectors to simultaneously increase or decrease. In this model, we assume if labeller's scores are very similar and consistent with the other labellers' scores or the average rating scores, the labeller might provide a more reliable credibility scores.

Pairwise rating similarity: It measures the pairwise rating value similarity between every pair of labellers. It shows to what extent labeller l_1 and labeller l_2 have cast similar consistence rating scores to every tweet message in the dataset. For each labeller, similarity value against other labellers is computed and then averaged. If similarity is calculated for every pair of labellers, a square symmetric matrix is formed that is equivalent to its transpose and can be used to identify the most similar labellers.

Average rating similarity: It computes the similarity between the labellers' rating scores and the average rating. In this scenario, the average rating of credibility scores for each tweet message is treated as the standard when evaluating the reliability of the labellers. Hence, labellers' reliability is measured by computing the similarity between their rating scores and the average rating.

4.1.1.1 Cosine similarity

A basic similarity function to measure similarity between two vectors is the unbounded inner product. If labeller l_1 score values tend to be high where labeller l_2 values are also high and

conversely low where l_2 score values are low, the inner product will be high which means the vectors are more similar.

$$inner\_product\ (l_1, l_2) = \sum_{i=1}^{n} CR[t_i, l_1]\,.\,CR[t_i, l_2]$$

$$inner\_product\ \ (l_1, AVGT) = \sum_{i=1}^{n} CR[t_i, l_1]\,.\,AVGT[t_i]$$

Cosine similarity model is a well-known model [78] for finding similarities between labellers' values. It is a normalized inner product - bounded between 0 and 1 if both vectors are non-negative, as in our dataset case. The value 1 means completely same and 0 means completely different. To calculate the similarity between labeller l_1 rating scores and other labellers' credibility scores l_j , the following formulas are used:

$$W_{l_1} = \frac{1}{m}\sum_{j=1}^{m} Cosine\,(l_1, l_j)$$

$$Cosine\,(l_1, l_j) = \frac{\sum_{i=1}^{n} CR[t_i, l_1]\,.\,CR[t_i, l_j]}{\sqrt{\sum_{i=1}^{n} CR[t_i, l_1]^2}\,.\sqrt{\sum_{i=1}^{n} CR[t_i, l_j]^2}}$$

Where W_{l_1}, the weighted value of labeller l_1 , is the average similarity values against all the other labellers. To compute the similarity between labeller l_1 values and the average credibility rating vector $AVGT$, the following formula is used:

$$W_{l_1} = \frac{\sum_{i=1}^{n} CR[t_i, l_1]\,.\,AVGT[t_i]}{\sqrt{\sum_{i=1}^{n} CR[t_i, l_1]^2}\,.\sqrt{\sum_{i=1}^{n} AVGT[t_i]^2}}$$

The table that displays the computed weights for all labellers after applying both pairwise rating similarity and average rating similarity is listed in Appendix B (Table B-1, Table B-2). A comparison between the resulted tweet messages' labels $TLabel_{t_i}$ and experts rating values is listed below in Table 4-3 using Krippendorff's alpha agreement [79]. Measures such as Pearson correlation coefficient (PCC), and Intraclass correlation coefficient (ICC) - single lower are also considered to reveal a more detailed quantification of the agreement. With reference to the table below, the average rating similarity denotes the highest agreement, correlation level and best overlap from both measures.

Table 4-3 Labelling after applying Cosine similarity compared to experts' labelling

Cosine Similarity		Expert1_A	Expert2_N	Expert3_E	Average
PCC	Pairwise rating similarity	0.4939	0.1772	0.3561	0.3424
	Average rating similarity	**0.6933**	0.3401	0.3711	0.4682
Alpha Agreement	Pairwise rating similarity	0.5524	0.0527	0.2219	0.2757
	Average rating similarity	**0.7120**	0.1020	0.2912	0.3684
ICC	Pairwise rating similarity	0.4940	0.1110	0.2950	0.3000
	Average rating similarity	**0.6790**	0.1910	0.3280	0.3993

4.1.1.2 Pearson similarity

Pearson's correlation coefficient is a proficient technique to measure the rating similarity between the labellers. It is the most conventional and common method [80] adopted by the scientific community to represent labellers' reliability. Pearson's correlation coefficient is the cosine similarity between average-cantered versions of two vectors and bounded between -1 and 1. To calculate the similarity between labeller l_1 and another labeller l_j , following formula is used:

$$Pearson(l_1, l_j) = \frac{\sum_{i=1}^{n}\left(CR[t_i, l_1] - \overline{CR[t_i, l_1]}\right).\left(CR[t_i, l_j] - \overline{CR[t_i, l_j]}\right)}{\sqrt{\sum_{i=1}^{n}(CR[t_i, l_1] - \overline{CR[t_i, l_1]})^2}.\sqrt{\sum_{i=1}^{n}(CR[t_i, l_j] - \overline{CR[t_i, l_j]})^2}}$$

To find the estimated weight for labeller l_1, similarity values against all other labellers are computed and then averaged:

$$W_{l_1} = \frac{1}{m}\sum_{j=1}^{m} Pearson\ (l_1, l_j)$$

For the average rating similarity, the Pearson correlation coefficient between the rating vector of labeller l_1 and the corresponding tweets' average rating vector $AVGT$ is given by:

$$W_{l_1} = \frac{\sum_{i=1}^{n}\left(CR[t_i, l_1] - \overline{CR[t_i, l_1]}\right).\left(AVGT[t_i] - \overline{AVGT[t_i]}\right)}{\sqrt{\sum_{i=1}^{n}\left(CR[t_i, l_1] - \overline{CR[t_i, l_1]}\right)^2}.\sqrt{\sum_{i=1}^{n}\left(AVGT[t_i] - \overline{AVGT[t_i]}\right)^2}}$$

As in previous methods, the tables illuminate the computed weights for all labellers after applying both pairwise rating similarity and average rating similarity using Pearson correlation coefficient is listed in Appendix B (Table B-1, Table B-2). A validation of this model is shown in Table 4-4 where the resulted tweet messages' labels $TLabel_{t_i}$ is compared to experts rating values. It emphasizes that average rating similarity surpasses pairwise similarity to estimate labels that agree and correlate with experts' labelling.

Table 4-4 Labelling after applying PCC similarity compared to experts' labelling

PCC Similarity		Expert1_A	Expert2_N	Expert3_E	Average
PCC	Pairwise rating similarity	0.5759	0.1862	0.3162	0.3594
	Average rating similarity	**0.6567**	0.3266	0.4078	0.4637
Alpha Agreement	Pairwise rating similarity	0.6395	0.0570	0.1820	0.2928
	Average rating similarity	**0.6707**	0.0767	0.3369	0.3614
ICC	Pairwise rating similarity	0.5770	0.1190	0.2580	0.3180
	Average rating similarity	**0.6380**	0.1770	0.3670	0.3940

4.1.1.3 Jaccard similarity

Tanimoto coefficient, known as the extended Jaccard coefficient, is another measurement for comparing the similarity and diversity of two vectors. It is a bounded measure between 0 and 1 used to compute the extent to which labellers are similar in their credibility judgment scores. To calculate the pairwise Jaccard, the similarity value between labeller l_1 and another labeller l_j is calculated and then similarity values against all the other labellers is also computed and averaged as shown below:

$$W_{l_1} = \frac{1}{m}\sum_{j=1}^{m} Jaccard\,(l_1, l_j)$$

$$Jaccard\,(l_1, l_j) = \frac{\sum_{i=1}^{n} CR[t_i, l_1]\,.\,CR[t_i, l_j]}{\sum_{i=1}^{n} CR[t_i, l_1]^2 + \sum_{i=1}^{n} CR[t_i, l_j]^2 - \sum_{i=1}^{n} CR[t_i, l_1]\,.\,CR[t_i, l_j]}$$

To calculate the similarity between l_1 values and the average rating $AVGT$, this following formula is used:

$$W_{l_1} = \frac{\sum_{i=1}^{n} CR[t_i, l_1]\,.\,AVGT[t_i]}{\sum_{i=1}^{n} CR[t_i, l_1]^2 + \sum_{i=1}^{n} AVGT[t_i]^2 - \sum_{i=1}^{n} CR[t_i, l_1]\,.\,AVGT[t_i]}$$

The labellers' weights are listed in Table B-1 and Table B-2 in Appendix B where a comparison between the constructed tweet messages' labels $TLabel_{t_i}$ with experts rating values is shown below in Table 4-5. As presented in the table below, the average rating similarity clearly shows the highest agreement and correlation level with experts' labelling.

Table 4-5 Labelling after applying Jaccard similarity compared to experts' labelling

Jaccard Similarity		Expert1_A	Expert2_N	Expert3_E	Average
PCC	Pairwise rating similarity	0.4649	0.1862	0.3670	0.3393
	Average rating similarity	**0.6349**	0.3204	0.3836	0.4463
Alpha Agreement	Pairwise rating similarity	0.5416	0.0570	0.2185	0.2724
	Average rating similarity	**0.6290**	0.0388	0.3329	0.3336
ICC	Pairwise rating similarity	0.4660	0.1190	0.2990	0.2947
	Average rating similarity	**0.6130**	0.1690	0.3510	0.3777

4.1.1.4 Intra-class similarity

Intraclass correlation coefficient (ICC) could also be used as measure of reliability as it calculates the correlations between pairs of vectors, taking into account the variance between the rating values. ICC similarity between the average vector and labellers' ratings is calculated in this section to assess the degree that labellers provided consistency in their ratings with the average rating. The labellers' weights listed in Table B-2 in Appendix B are evaluated using SPSS tool[12] with a two-way mixed, absolute agreement, and single-measures ICC. Cicchetti & Sparrow 1981 [81] provide classified levels of inter-labeller ICC values as follows: "poor" for ICC values less than 0.40, "fair" for values between 0.40 and 0.59, "good" for values between 0.60 and 0.74, and "excellent" for values between 0.75 and 1.0. A comparison between labels obtained by experts and labels constructed using ICC similarity method is provided below in Table 4-6.

Table 4-6 Labelling after applying ICC similarity compared to experts' labelling

ICC Similarity		Expert1_A	Expert2_N	Expert3_E	Average
PCC	Average rating similarity	0.6616	0.3110	0.4670	0.4799
Alpha Agreement	Average rating similarity	**0.6696**	0.0336	0.4103	0.3712
ICC	Average rating similarity	0.6340	0.1610	0.4300	0.4083

4.1.2 Weighted labellers algorithm using similarity model

The following proposed algorithm was applied using previous measures to compute stable labellers' weights, and then estimate the correct labels of tweet messages. This algorithm iteratively updates initial labellers' weights on each round until the weight values reach a stable point. The key point concerning this algorithm is that a labeller who has more similar credibility values to the weighted average values should have a higher reliability weight. A model used Pearson's correlation coefficient to determine users reputations has been proposed by Zhou et al. 2011 [82]. In their algorithm, all users with negative correlations values were assigned to zero reputations' weights and then the weighted average were recalculated. However, our method included all labellers' weights to compute the weighted average, plus they received weights depending on their correlation values rather than discounting their contributions.

[12] IBM SPSS Statistics 22 Documentation,
http://www-01.ibm.com/support/docview.wss?uid=swg27038407

Detailed steps used to compute the reliability weight of a labeller l_j are shown below:

---

1. *Compute the initial similarity weight of labeller l_j; which is calculated previously using the similarity model.*

$$WCosine_{l_j} = \frac{\sum_{i=1}^{n} CR[t_i, l_j] \,.\, AVGT[t_i]}{\sqrt{\sum_{i=1}^{n} CR[t_i, l_j]^2} \,.\, \sqrt{\sum_{i=1}^{n} AVGT[t_i]^2}}$$

$$WPearson_{l_j} = \frac{\sum_{i=1}^{n} \left(CR[t_i, l_j] - \overline{CR[t_i, l_j]}\right).\left(AVGT[t_i] - \overline{AVGT[t_i]}\right)}{\sqrt{\sum_{i=1}^{n} (CR[t_i, l_j] - \overline{CR[t_i, l_j]})^2} \,.\sqrt{\sum_{i=1}^{n} (AVGT[t_i] - \overline{AVGT[t_i]})^2}}$$

$$WJaccard_{l_j} = \frac{\sum_{i=1}^{n} CR[t_i, l_j] \,.AVGT[t_i]}{\sum_{i=1}^{n} CR[t_i, l_j]^2 + \sum_{i=1}^{n} AVGT[t_i]^2 - \sum_{i=1}^{n} CR[t_i, l_j] \,.AVGT[t_i]}$$

2. *After computing labeller's weight using each measure, we standardize the range of weight values and normalize it between 0 and 1 using min-max normalization.*
3. *Then, the weighted average of rating for the tweet t_i is computed (weighted average equals the aggregated rating from the subset of labellers rated tweet t_i multiplied by their updated weight and divided by the total labellers' weight).*

$$W\_AVGT\_Cosine[t_i] = \frac{\sum_{j=1}^{m_{t_i}} WCosine_{l_j} . CR[t_i, l_j]}{\sum_{j=1}^{m_{t_i}} WCosine_{l_j}}$$

$$W\_AVGT\_Pearson[t_i] = \frac{\sum_{j=1}^{m_{t_i}} WPearson_{l_j} . CR[t_i, l_j]}{\sum_{j=1}^{m_{t_i}} WPearson_{l_j}}$$

$$W\_AVGT\_Jaccard[t_i] = \frac{\sum_{j=1}^{m_{t_i}} WJaccard_{l_j} . CR[t_i, l_j]}{\sum_{j=1}^{m_{t_i}} WJaccard_{l_j}}$$

4. *Apply the selected similarity measure between the rating vector of labeller l_j and the corresponding updated tweets' average weighted vector.*

$$WCosine_{l_j} = \frac{\sum_{i=1}^{n} CR[t_i, l_j] \,.W\_AVGT\_Cosine[t_i]}{\sqrt{\sum_{i=1}^{n} CR[t_i, l_j]^2} \,.\, \sqrt{\sum_{i=1}^{n} W\_AVGT\_Cosine[t_i]^2}}$$

$$WPearson_{l_j} = \frac{\sum_{i=1}^{n} \left(CR[t_i, l_j] - \overline{CR[t_i, l_j]}\right) .\left(W\_AVGT\_Pearson[t_i] - \overline{W\_AVGT\_Pearson[t_i]}\right)}{\sqrt{\sum_{i=1}^{n} \left(CR[t_i, l_j] - \overline{CR[t_i, l_j]}\right)^2} \,.\sqrt{\sum_{i=1}^{n} \left(W\_AVGT\_Pearson[t_i] - \overline{W\_AVGT\_Pearson[t_i]}\right)^2}}$$

$$WJaccard_{l_j} = \frac{\sum_{i=1}^{n} CR[t_i, l_j] \,.\, W\_AVG\_Jaccard[t_i]}{\sum_{i=1}^{n} CR[t_i, l_j]^2 + \sum_{i=1}^{n} W\_AVG\_Jaccard[t_i]^2 - \sum_{i=1}^{n} CR[t_i, l_j] \,.\, W\_AVG\_Jaccard[t_i]}$$

5. *Iterate 2, 3, and 4 until the average difference of labellers' weight between iteration is less than a threshold 10^{-5} and the rank of labellers' weights remains stable for at least 5 continuous iterations.*

$$\frac{1}{m_{t_i}} \sum_{j=1}^{m_{t_i}} \left| WCosine_{l_j} - \; WCosine_{l_j}{}' \right| < 10^{-5}$$

$$\frac{1}{m_{t_i}} \sum_{j=1}^{m_{t_i}} \left| WPearson_{l_j} - \; WPearson_{l_j}{}' \right| < 10^{-5}$$

$$\frac{1}{m_{t_i}} \sum_{j=1}^{m_{t_i}} \left| WJaccard_{l_j} - \; WJaccarad_{l_j}{}' \right| < 10^{-5}$$

The labellers' weights after applying the iterative algorithm using similarity model is displayed in Table B-3, Appendix B. A comparison between the constructed tweet messages' labels $TLabel_{t_i}$ using similarity model before and after applying proposed algorithm, and experts rating values is listed below in Table 4-7 using Krippendorff's alpha agreement. Meanwhile since labellers' similarity weighting values resulted after the cosine similarity algorithm has changed slightly, the labelling agreement values with experts remain the same.

Table 4-7 Labelling after applying similarity algorithms compared to experts' labelling

Similarity algorithms - Krippendorff's Alpha		**Expert1_A**	**Expert2_N**	**Expert3_E**	**Average**
Cosine Similarity	Average rating similarity	**0.7120**	0.1020	0.2912	0.3684
	Average rating similarity algorithm	**0.7120**	0.1020	0.2912	0.3684
Pearson Similarity	Average rating similarity	**0.6707**	0.0767	0.3369	0.3614
	Average rating similarity algorithm	**0.6096**	-0.0112	0.3496	0.3160
Jaccard Similarity	Average rating similarity	**0.6290**	0.0388	0.3329	0.3336
	Average rating similarity algorithm	**0.5974**	0.0335	0.2685	0.2998

4.1.3 Accuracy model

In addition to similarity measures, we also measured labellers' rating accuracy in order to identify to what distance each labeller cast rating values as other labellers' values and as well as average rating values. In this model, we assume if the differences between a labeller's ratings and other labellers' values as well as the average rating scores are small, the labeller might provide a more accurate credibility scores. To estimate the labellers' distance accuracies, we computed the rating differences between labellers' ratings and average rating values using different methods. A pairwise rating accuracy is used to compute the differences between each labeller assigned credibility values and other labellers' values using the whole tweets' messages. In addition, average rating accuracy is used to calculate the differences between the labeller rating values and the average rating values. Labellers' weights were assigned according to the magnitude of their computed differences, translating to the greater the difference, the smaller the weight.

4.1.3.1 Pairwise rating differences accuracy

It measures the distance between a labeller assigned credibility values and other labellers' values. To calculate this distance for each labeller, the difference rating values against other labellers is computed and then averaged. The accuracy formula assumes that if the difference

of labeller's ratings with other labeller's ratings is minimal, then the participant receives correct rating values. To compute the weight W_{l_1} for labeller l_1 using pairwise rating accuracy, the following formula is used:

$$W_{l_1} = 1 - \frac{\sum_{i=1}^{n_{l_1}} \sum_{j=1}^{m_{t_i}} |CR[t_i, l_1] - CR[t_i, l_j]|}{\sum_{i=1}^{n_k} \sum_{j=1}^{m_l} 1}$$

The computed weights for all labellers after applying pairwise rating accuracy is listed in Table B-4, Appendix B. A comparison between the derived tweet messages' labels $TLabel_{t_i}$ with experts' rating values is shown below in Table 4-8. As illustrated in the table below, pairwise rating accuracy did not result in better agreements or correlations compared to experts' labelling. This outcome was also the case with the similarity model where similarity using average rating was better than pairwise similarity in order to construct labelling match the experts labelling.

Table 4-8 Labelling after applying pairwise rating differences accuracy compared to experts' labelling

Pairwise differences accuracy	Expert1_A	Expert2_N	Expert3_E	Average
PCC	0.5071	0.2008	0.3291	0.3457
Alpha Agreement	**0.5831**	0.0814	0.1674	0.2773
ICC	0.5090	0.1330	0.2600	0.3007

4.1.3.2 Average rating accuracy

To estimate to what extent labellers' credibility scores correspond to the average ratings, different methods based on computing the rating scores dispersion (also called variability, or spread) have been proposed below:

4.1.3.2.1 Average absolute deviation

Accuracy is computed by identifying the difference between the labellers' rating values and the average rating which represents the deviation of this labeller's rating scores. Labellers' weights are calculated simply by adding up the deviation of each score from the average rating and then dividing by the number of credibility scores. The smallest quantity is the difference, the greater one is the weight. To compute the weight W_{l_1} for labeller l_1 , the following formula is used:

$$W_{l_1} = 1 - \frac{\sum_{i=1}^{n_{l_1}} |CR[t_i, l_1] - AVGT[t_i]|}{\sum_{i=1}^{n_{l_1}} 1}$$

Table displays the computed weights for all labellers after applying average absolute deviation is listed in Table B-5, Appendix B. A comparison between the derived tweet messages' labels $TLabel_{t_i}$ with experts' rating values is shown below in Table 4-9. With reference to the table below, average rating accuracy produces labelling that has a higher agreement and correlation values with experts' labelling in comparison to the previous pairwise accuracy method.

Table 4-9 Labelling after applying average absolute deviation accuracy compared to experts' labelling

Average absolute deviation accuracy	Expert1_A	Expert2_N	Expert3_E	Average
PCC	0.5618	0.1644	0.3444	0.3569
Alpha Agreement	**0.6075**	0.0294	0.2343	0.2904
ICC	0.5590	0.0990	0.2920	0.3167

4.1.3.2.2 Normalized deviation

This method is proposed by Ignjatovic et al. 2009 [83] and it is extended version of the previous average absolute deviation technique. It requires computing the normalized average labeller's weight using previous method according to the weights of other labellers, so the weight assigned to each labeller is related to the total labellers' deviations. To compute the weight W_{l_1} for labeller l_1 , the following formula has been used:

$$Avg\_abs\_dev(l_1, AVGT) = \frac{\sum_{i=1}^{n_{l_1}} |CR[t_i, l_1] - AVGT[t_i]|}{\sum_{i=1}^{n_{l_1}} 1}$$

$$W_{l_1} = \frac{1 - \frac{Avg\_abs\_dev(l_1, AVGT)}{\sum_{j=1}^{m_{t_i}} Avg\_abs\_dev(l_j, AVGT)}}{\sum_{i=1}^{m_{t_i}} 1 - \frac{Avg\_abs\_dev(l_i, AVGT)}{\sum_{j=1}^{m_{t_i}} Avg\_abs\_dev(l_j, AVGT)}}$$

Ignjatovic et al. 2009 [83] proposed using the previous formula to update labellers weights. Steps of the applied algorithm are shown below (with slight modification):

---

1. *Compute the initial normalized deviation weight for labeller l_j using the following formula:*

$$Avg\_abs\_dev(l_j, AVGT) = \frac{\sum_{i=1}^{n_{l_j}} |CR[t_i, l_j] - AVGT[t_i]|}{\sum_{i=1}^{n_{l_j}} 1}$$

$$W_{l_j} = \frac{1 - \frac{Avg\_abs\_dev(l_j, AVGT)}{\sum_{k=1}^{m_{t_i}} Avg\_abs\_dev(l_k, AVGT)}}{\sum_{i=1}^{m_{t_i}} 1 - \frac{Avg\_abs\_dev(l_i, AVGT)}{\sum_{k=1}^{m_{t_i}} Avg\_abs\_dev(l_k, AVGT)}}$$

2. *Find the weighted average of rating for the tweet t_i which equals the aggregated rating from the subset of labellers rated tweet t_i multiplied by their initial/updated weight and divided by the total labellers' weights.*

$$W\_AVGT\_ND[t_i] = \frac{\sum_{j=1}^{m_{t_i}} W_{l_j} . CR[t_i, l_j]}{\sum_{j=1}^{m_{t_i}} W_{l_j}}$$

3. *Apply the normalized deviation measure between the rating vector of labeller l_j and the corresponding updated tweets' average weighted vector:*

$$Avg\_abs\_dev(l_j, W\_AVGT\_ND) = \frac{\sum_{i=1}^{n_{l_1}} |CR[t_i, l_j] - W\_AVGT\_ND[t_i]|}{\sum_{i=1}^{n_{l_j}} 1}$$

$$W_{l_j} = \frac{1 - \frac{Avg\_abs\_dev(l_j, W\_AVGT\_ND)}{\sum_{k=1}^{m_{t_i}} Avg\_abs\_dev(l_k, W\_AVGT\_ND)}}{\sum_{i=1}^{m_{t_i}} 1 - \frac{Avg\_abs\_dev(l_i, W\_AVGT\_ND)}{\sum_{k=1}^{m_{t_i}} Avg\_abs\_dev(l_k W\_AVGT\_ND)}}$$

4. *Iterate 2, and 3 until the average difference of labellers' weights between iteration is less than a threshold 10^{-5} and the rank of labellers' weights remains stable for at least 5 continuous iterations.*

$$\frac{1}{m_{t_i}} \sum_{j=1}^{m_{t_i}} \left| W_{l_j} - W_{l_j}' \right| < 10^{-5}$$

Although labellers' weights has been updated using algorithm above, the credibility labelling constructed using the computed labellers' weights remained the same as before applying the algorithm since the weights of labellers has changed slightly. A comparison between the derived tweet messages' labels $TLabel_{t_i}$ with experts' rating values is shown below in Table 4-10. A list of the computed labellers' weights using normalized deviation and normalized deviation algorithm is found in Table B-5, Appendix B.

Table 4-10 Labelling after applying normalized average absolute deviation accuracy algorithm compared to experts' labelling

Normalized average absolute deviation	Expert1_A	Expert2_N	Expert3_E	Average
PCC	0.5618	0.1644	0.3444	0.3569
Alpha Agreement	**0.6075**	0.0294	0.2343	0.2904
ICC	0.5590	0.0990	0.2920	0.3167

4.1.3.2.3 Variance

Another method for calculating the deviation of a labeller's credibility scores from the average rating scores is the use of variance. To compute the variance, squared deviations are totalled and then averaged by the number of credibility scores. Then the accuracy weight for each labeller is estimated by averaging the difference between the aggregated values for all labellers' deviations and each labeller's deviation [84].

$$Deviation_{l_1} = \frac{\sum_{i=1}^{n_{l_1}} (CR[t_i, l_1]) - AVGT[t_i])^2}{\sum_{i=1}^{n_{l_1}} 1}$$

$$Labellers\_avg = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_{t_i}} 1}{n}$$

$$W_{l_1} = \frac{(\sum_{j=1}^{m} Deviation_{l_j}) - Deviation_{l_1}}{Labellers\_avg \,.\sum_{j=1}^{m} Deviation_{l_j}}$$

Table B-6 displays the computed weights for all labellers after applying this method is listed in Appendix B. A comparison between derived tweet messages' labels $TLabel_{t_i}$ with experts' rating values is presented below in Table 4-11.

Table 4-11 Labelling after applying variance accuracy compared to experts' labelling

Variance accuracy	Expert1_A	Expert2_N	Expert3_E	Average
PCC	0.6024	0.1772	0.3066	0.3620
Alpha Agreement	**0.6489**	0.0527	0.1857	0.2958
ICC	0.6010	0.1110	0.2540	0.3220

4.1.3.2.4 Variance by topic

Another way to compute labellers' accuracy is to reflect upon the topic type while computing the average deviation. The previous average method is based on assigning uniform weights to all tweet messages for all topics to assess a labeller's overall weight. In this section, we calculated a different weight for each topic and the respective average of these weights. The same steps (as in previous method) are followed including an additional step to average the weights for all topics.

$$Deviation_{l_1 p_1} = \frac{\sum_{i=1}^{n_{l_1 p_1}} (CR[t_i, l_1]) - AVGT[t_i])^2}{\sum_{i=1}^{n_{l_1 p_1}} 1}$$

$$WTopic_{l_1} = \frac{(\sum_{j=1}^{m_{p_1}} Deviation_{l_j}) - Deviation_{l_1}}{((\sum_{i=1}^{m_{p_1}} 1) - 1) \,.\sum_{j=1}^{m_{p_1}} Deviation_{l_j}}$$

$$W_{l_1} = \frac{\sum_{j=1}^{p_{l_1}} WTopic_{l_j}}{p_{l_1}}$$

The computed weights for all labellers after applying this method is listed in Table B-6, Appendix B. A comparison between derived tweet messages' labels $TLabel_{t_i}$ with experts' rating values is provided below in Table 4-12.

Table 4-12 Labelling after applying variance by topic accuracy compared to experts' labelling

Variance by topic accuracy	Expert1_A	Expert2_N	Expert3_E	Average
PCC	0.6024	0.1772	0.3066	0.3620
Alpha Agreement	**0.6489**	0.0527	0.1857	0.2958
ICC	0.6010	0.1110	0.2540	0.3220

4.1.3.2.5 Standard deviation

With this component, we assign weights to labellers according to the magnitude of their deviation; if labeller's credibility scores fall within the 1-standard deviation of the average, then labeller would be weighted more than others [85]. The following formulas are used for this:

$$Acc_{SD}(l_j) = \begin{cases} 1 & if\ (AVGT[t_i] - SDT[t_i]) \leq CR[t_i, l_j] \leq (AVGT[t_i] + SDT[t_i]) \\ \dfrac{(AVGT[t_i] - SDT[t_i]) - CR[t_i, l_j]}{ClassNo} & if\ CR[t_i, l_j] < (AVGT[t_i] - SDT[t_i]) \\ \dfrac{CR[t_i, l_j] - (AVGT[t_i] + SDT[t_i])}{ClassNo} & otherwise \end{cases}$$

$$W_{l_j} = \frac{1}{n_{l_j}} \sum_{j=1}^{n_{l_j}} Acc_{SD}(l_j)$$

Minor adjustments considered to this method where only the rating scores, which fall in 1-standard deviation range, are solely considered to calculate the weight. We calculated the labellers' weights using formulas below and study the difference between both methods:

$$Acc_{SD2}(l_j) = \begin{cases} 1 & if\ (AVGT[t_i] - SDT[t_i]) \leq CR[t_i, l_j] \leq (AVGT[t_i] + SDT[t_i]) \\ 0 & otherwise \end{cases}$$

$$W_{l_j} = \frac{1}{n_{l_j}} \sum_{j=1}^{n_{l_j}} Acc_{SD2}(l_j)$$

Table B-6 displays the computed weights for all labellers after applying both discussed methods is listed in Appendix B. A comparison between derived tweet messages' labels $TLabel_{t_i}$ with experts' rating values is shown below in Table 4-13. Both methods yielded the same labelling results compared to experts' labelling.

Table 4-13 Labelling after applying standard deviation range accuracy compared to experts' labelling

Standard deviation range accuracy	Expert1_A	Expert2_N	Expert3_E	Average
PCC	0.6024	0.1772	0.3066	0.3620
Alpha Agreement	**0.6489**	0.0527	0.1857	0.2958
ICC	0.6010	0.1110	0.2540	0.3220

4.1.4 Agreement model

This model is used to estimate the agreement ratio for every labeller with other labellers. In this study, we argue that a labeller's agreement or disagreement with other labellers reflect their overall reliability. We assume that a labeller who agrees more with other labellers makes more objective judgements. In line with this assumption, we computed the agreement using both

modified percentage agreement and alpha agreement. In both methods, we computed the agreement values for each pair of labellers and then averaged the resulted agreement values for each labeller paired in turn with the other labellers.

1. **Modified agreement percentage:** This is simply the proportion of tweet messages on which a labeller agrees on its credibility values with other labellers divided by the number of total judgements (agreements + disagreements). We calculated the agreement percentage of labeller l_1 applying the following formula:

$$Agree\left((l_1, l_j), t_i\right) = \begin{cases} 1 & if\ CR[t_i, l_1] = CR[t_i, l_j] \\ 0 & therwise \end{cases}$$

$$W_{l_1} = \frac{1}{m_{t_i}.n_{l_1}} \sum_{i=1}^{m_{t_i}} \sum_{j=1}^{n_{l_1}} Agree\left((l_1, l_j), t_i\right)$$

2. **Alpha agreement:** It is also used to compute the agreement for each labeller and others, in addition to calculating the averaged result. If alpha agreement is calculated for every pair of labellers, we can create a square symmetric matrix that is equal to its transpose. To calculate the alpha agreement for labeller l_1 , we used the following formula:

$$W_{l_1} = \frac{1}{m} \sum_{j=1}^{m} Alpha\left(l_1, l_j\right)$$

Table B-7 in Appendix B lists the computed weights for all labellers after applying both modified agreement percentage and alpha agreement. A comparison between derived tweet messages' labels $TLabel_{t_i}$ with experts' rating values using both agreement methods is shown below in Table 4-14.

Table 4-14 Labelling after applying agreement model compared to experts' labelling

Agreement model		Expert1_A	Expert2_N	Expert3_E	Average
PCC	Modified agreement percentage	0.5226	0.2058	0.3056	0.3447
Alpha Agreement		**0.6079**	0.1024	0.1351	0.2818
ICC		0.5230	0.1390	0.2390	0.3003
PCC	Alpha agreement	0.5950	0.3235	0.4536	0.4574
Alpha Agreement		**0.6200**	0.0576	0.3830	0.3535
ICC		0.5770	0.1730	0.4120	0.3873

4.1.5 Majority consensus model

This model is based on favouring labellers who generally conform to the popular community consensus. We proposed to measure the majority consensus of the labellers as it was an indicator for how close the labeller's credibility judgements corresponded to community

consensus. In this model, the study assumes that a labeller is more reliable and representative of the majority if his/her credibility judgement scores more closely endorse the ratings of the majority respondents. Consequently, a significant difference between the majority rating scores and labeller's credibility scores would trigger a lower labeller's weight. In line with this assumption, different methods have been applied below to drive the labellers' weights based on majority consensus model. A complete list of the computed labellers' weights after applying the methods and algorithms below is shown in Table B-8, Appendix B.

1. **Exact class matching:** To estimate the labellers' consensus ratio, we identified the majority class for each tweet. Then, we tabulated how many times the labeller assigned credibility score analogous to the majority rating score. To calculate the weight for labeller l_1 using exact class matching, we applied the following formula:

$$MajCons\,(l_1, t_i) = \begin{cases} 1 & if\ MajClass_{t_i} = CR[t_i, l_1] \\ 0 & otherwise \end{cases}$$

$$W_{l_1} = \frac{1}{n_{l_1}} \sum_{i=1}^{n_{l_1}} MajCons\,(l_1, t_i)$$

A comparison between derived tweet messages' labels $TLabel_{t_i}$ with experts' rating values is shown below in Table 4-15.

Table 4-15 Labelling after applying majority exact match compared to experts' labelling

Majority exact match	Expert1_A	Expert2_N	Expert3_E	Average
PCC	0.5226	0.2058	0.3056	0.3447
Alpha Agreement	**0.6079**	0.1024	0.1351	0.2818
ICC	0.5230	0.1390	0.2390	0.3003

2. **Class ratio:** Based on minimal modifications to the previous method, this technique focused on assigning a value for each labeller even if the score did not match the majority class and depended on the ratio value of the selected class. To estimate the labellers' consensus ratios, we calculated the credibility values' count $ClassC$ for each credibility class {1, 2, 3} and for each message tweet. Then we computed the weight for every labeller depending on the ratio of chosen class. The following formulas are used to evaluate the weight of labeller l_1 using this technique:

$$Ratio\,(t_i, ClassC_1) = \frac{ClassC_1}{\sum_{j=1}^{ClassNo} ClassC_j}$$

$$Ratio\,(t_i, ClassC_2) = \frac{ClassC_2}{\sum_{j=1}^{ClassNo} ClassC_j}$$

$$Ratio\,(t_i, ClassC_3) = \frac{ClassC_3}{\sum_{j=1}^{ClassNo} ClassC_j}$$

$$W_{l_1} = \frac{1}{n_{l_1}} \sum_{i=1}^{n_{l_1}} Ratio\left(t_i, ClassC_{CR[t_i,l_1]}\right)$$

A comparison between derived tweet messages' labels $TLabel_{t_i}$ with experts' rating values is shown below in Table 4-16.

Table 4-16 Labelling after applying majority class ratio compared to experts' labelling

Majority class ratio	Expert1_A	Expert2_N	Expert3_E	Average
PCC	0.5226	0.2058	0.3056	0.3447
Alpha Agreement	**0.6079**	0.1024	0.1351	0.2818
ICC	0.5230	0.1390	0.2390	0.3003

3. **Normalized class ratio:** It is the same method as the previous one but with an additional step to normalize the credibility values' count for each credibility class {1, 2, 3} where:

$$\sum_{j=1}^{ClassNo} ClassC_j{}^2 = 1$$

Then, calculate the ratio for each normalized credibility count value and for each tweet message, Where $ClassC_1$ is the count number for class {1}, $ClassC_2$ is the count number for class {2}, $ClassC_3$ is the count number for class {3}.

$$Ratio\ (t_i, ClassC_1) = \frac{ClassC_1}{\sqrt{\sum_{j=1}^{ClassNo} ClassC_j{}^2}}$$

$$Ratio\ (t_i, ClassC_2) = \frac{ClassC_2}{\sqrt{\sum_{j=1}^{ClassNo} ClassC_j{}^2}}$$

$$Ratio\ (t_i, ClassC_3) = \frac{ClassC_3}{\sqrt{\sum_{j=1}^{ClassNo} ClassC_j{}^2}}$$

To calculate the weight for each labeller, we aggregated the ratios of all credibility count classes for all tweet messages depending on the participants' assigned class for each tweet message. The weight for labeller W_{l_1} is as following:

$$W_{l_1} = \frac{1}{n_{l_1}} \sum_{i=1}^{n_{l_1}} Ratio\left(t_i, ClassC_{CR[t_i,l_1]}\right)$$

A comparison between derived tweet messages' labels $TLabel_{t_i}$ with experts' rating values is presented below in Table 4-17 .

Table 4-17 Labelling after applying majority normalized class ratio compared to experts' labelling

Majority normalized class ratio	Expert1_A	Expert2_N	Expert3_E	Average
PCC	0.5226	0.2058	0.3056	0.3447
Alpha Agreement	**0.6079**	0.1024	0.1351	0.2818
ICC	0.5230	0.1390	0.2390	0.3003

4.1.5.1 Weighted labellers algorithm using majority consensus model

By using the normalized class ratio technique, a modified algorithm proposed by Allahbakhsh and Ignjatovic 2015 [86] is applied to update weighting of the labellers till stable weights are obtained. Steps required for this algorithm is shown below:

---

1. *Set the initial weight for all labellers to 1.*
   $W_{l_j} = 1$
2. *Calculate the credibility values' count for each credibility class {1, 2, 3} and for each message tweet then normalize the credibility values' count where:*
   $$\sum_{j=1}^{ClassNo} ClassC_j{}^2 = 1$$
3. *Calculate the ratio for each credibility class and for each tweet message, where $ClassC_1$ is the count number for class {1} , $ClassC_2$ is the count number for class {2}, $ClassC_3$ is the count number for class {3}*
   $$Ratio\ (t_i, ClassC_1) = \frac{ClassC_1}{\sqrt{\sum_{j=1}^{ClassNo} ClassC_j{}^2}} = \frac{\sum_{j=1}^{m_l} W_{l_j} \forall\ CR[t_i, l_j] = 1}{\sqrt{\sum_{j=1}^{m_l} W_{l_j}{}^2}\ \forall\ CR[t_i, l_j]}$$
   $$Ratio\ (t_i, ClassC_2) = \frac{ClassC_2}{\sqrt{\sum_{j=1}^{ClassNo} ClassC_j{}^2}} = \frac{\sum_{j=1}^{m_l} W_{l_j} \forall\ CR[t_i, l_j] = 2}{\sqrt{\sum_{j=1}^{m_l} W_{l_j}{}^2}\ \forall\ CR[t_i, l_j]}$$
   $$Ratio\ (t_i, ClassC_3) = \frac{ClassC_3}{\sqrt{\sum_{j=1}^{ClassNo} ClassC_j{}^2}} = \frac{\sum_{j=1}^{m_l} W_{l_j} \forall\ CR[t_i, l_j] = 3}{\sqrt{\sum_{j=1}^{m_l} W_{l_j}{}^2}\ \forall\ CR[t_i, l_j]}$$
4. *Calculate the weight for each labeller by aggregating the ratio of all credibility count classes for all tweet messages depending on the class he/she assigned for each tweet message. The weight for labeller l_1 is as following:*
   $$W_{l_1} = \frac{1}{n_{l_1}} \sum_{i=1}^{n_{l_1}} Ratio\ (t_i, ClassC_{CR[t_i, l_1]})$$
5. *Repeat from step 3 by calculating the new ratio value for each credibility class using the new labellers' weights. Stop iterations when there is no significant labellers' weights difference means the average difference of labellers' weights between iterations is less than a threshold 10^{-5} and the rank of labellers' weights remain stable for at least 5 continues iterations.* $\frac{1}{m_{t_i}} \sum_{j=1}^{m_{t_i}} \left| W_{l_j} - W_{l_j}{}' \right| < 10^{-5}$

A comparison between derived tweet messages' labels $TLabel_{t_i}$ with experts rating values is shown below in Table 4-18.

Table 4-18 Labelling after applying majority normalized class ratio algorithm compared to experts' labelling

Majority normalized class ratio algorithm	Expert1_A	Expert2_N	Expert3_E	Average
PCC	0.5226	0.2058	0.3056	0.3447
Alpha Agreement	**0.6079**	0.1024	0.1351	0.2818
ICC	0.5230	0.1390	0.2390	0.3003

Labelling results compared before and after applying the algorithm are the same which means that labellers' weights were stable from the first iteration.

4.1.6 Propensity to trust model

In this study, we assumed that one of the factors which affected credibility judgments labelling is labellers' propensity to over-rate or under-rate tweet messages credibility. Propensity to trust is frequently understood as an individual's general willingness to trust others [87]. Labellers, who demonstrate a high propensity to trust, tend to respond more generously in their credibility evaluations. On the other hand, labellers who typically give low credibility scores tend to display a low propensity to trust. In the process of observing random selected credibility ratings from the dataset shown in Table 4-19, we noted some interesting findings: tweet message#1 generally received high ratings except from labeller#10. However, when we scrutinized ratings of labeller#10, we discovered that this participant expressed a harsher opinion, or in technical terminology – a lower propensity to trust on all the tweets, compared to the other labellers.

Table 4-19 Observations from the rating table

T/L	L#1	L#2	L#3	L#4	L#5	L#6	L#7	L#8	L#9	L#10	L#11	L#12	L#13	……
T#1	5	5	5	3	5	5	4	5	5	2	3	3	3	
T#2	1	2	1	3	4	3	2	1	5	1	2	2	1	
T#3	2	3	2	2	3	2	3	4	3					
T#4	4	3	2	2	2	4	2	4	4	1	2	3	2	
T#5	2	3	2	4	4	4	2	2	4	2	3	2	1	
T#6	3	3	1	4	3		2	1	5	1	3	2	1	
T#7	3	3	1	3	3	4	3	1	4	2	2	1	3	
T#8	3	3	1	4	2	3	2	1	2	1	1	1	1	
T#9	5	5	5	5	3	4	2	4	5	2	5	1	3	

The propensity to trust for a particular labeller is measured as an average credibility score given by that labeller relative to the average score across all the labellers. The basic idea is to identify labellers with a low propensity to trust. For example, if a labeller assigns positive credibility scores to all messages, their propensity to trust may be considered either very high, or compatible to the context, depending on the average scores specified by other labellers. We calculated the propensity to trust of labeller l_1 by the average difference between his/her rating scores and the average ratings for the same tweet messages [83]. Also, we computed the

percentage of times a labeller l_1 assigns low credibility class {1} from all his/her credibility judgements.

$$PropTrustsign(l_1, AVGT) = \frac{\sum_{i=1}^{n_{l_1}} (CR[t_i, l_1] - AVGT[t_i])}{\sum_{i=1}^{n_{l_1}} 1}$$

$$PropTrust(l_1, AVGT) = \frac{\sum_{i=1}^{n_{l_1}} |CR[t_i, l_1] - AVGT[t_i]|}{\sum_{i=1}^{n_{l_1}} 1}$$

$$Perc(l_1) = \frac{\sum_{i=1}^{n_{l_1}} 1 \; \forall \; CR[t_i, l_1] = 1}{\sum_{i=1}^{n_{l_1}} 1}$$

$$Propensity\ to\ trust\ (l_1) = \begin{Bmatrix} Low & if\ PropTrustsign(l_1, AVGT) + PropTrust(l_1, AVGT) \leq 0.1 \\ Normal & Otherwise \end{Bmatrix}$$

As the used dataset inclined to low credibility class, the high propensity to trust is not considered in this study. For identifying labellers with a low propensity to trust, we checked all $PropTrust$, $PropTrustsign$, and $Perc$ values for each labeller. With a large value of $PropTrust$, a large negative value of $PropTrustsign$, and a large $Perc$ then labeller l_1 considered as a harsh labeller with a low propensity of trust [83]. We categorized all labellers less than or equal a suggested threshold of 0.1 with a low propensity to trust; the suggested threshold will guarantee to include all labellers who have large negative values of the average deviation ($PropTrustsign \leq 0.5$). Table 4-20 records the labellers identified with low propensity to trust using $PropTrust$, $PropTrustsign$, and $Perc$ measures. A full list of the computed labellers' average deviations for detecting labellers' propensity to trust is shown in Table B-9, Appendix B.

Table 4-20 Labellers with low propensity to trust

Judges	PropTrust (average absolute deviation)	PropTrustsign (average deviation)	Perc Class {1} %	PropTrustsign + PropTrust	Rank
Labeller10	0.8121	-0.7525	90.53 %	0.0597	3
Labeller13	0.7361	-0.6309	82.76 %	0.1051	7
Labeller20	0.7773	-0.6966	86.39 %	0.0807	4
Labeller21	0.6901	-0.6060	78.16 %	0.0841	5
Labeller24	0.6121	-0.5642	89.47 %	0.0478	1
Labeller26	0.7815	-0.6973	93.88 %	0.0843	6
Labeller29	0.6692	-0.6103	94.44 %	0.0589	2

Later on, after applying the proposed labellers' weights aggregation model, all the labellers who were identified with low propensity to trust were not from the top reliable labellers. Three were from the moderate-reliability list whereas the others were found in the low-reliability level list. It is worth mentioning that expert **Expert2_N** who has the lowest agreement value with the

crowd compared to other experts also reports a very low propensity to trust with almost 95% of his/her ratings within a low-credibility class.

4.2 Conclusions from Labellers' Evaluation and Weighting

In this chapter we introduced an extra step prior to the classifier building to solve the problem of labelling disagreements between labellers which was overlooked in other studies. We introduced a framework encompassed different measurements for evaluating labellers' weights and used experiments to assess how the proposed techniques can enhance the fairness and quality of the credibility labelling. We validated proposed measurements by comparing the resulted labelling obtained after applying the framework of measurements with experts rating values. Table 4-21 below summarizes the agreement values between the ground truth vectors after applying the measurement and **Expert1_A** ratings using Krippendorff's alpha compared to the labelling assembled using majority voting options: Maj_Class2, Maj_Low, Maj_Hi, Maj_N (formulated previously in chapter 3).

Table 4-21 Labelling after applying proposed measures compared to experts' labelling

Labelling using simple majority voting compared to experts' labelling	**Expert1_A**
Maj_Class2	0.6055
Maj_Low	0.5831
Maj_Hi	0.6075
Maj_N	0.6055
Labelling after applying similarity measures compared to experts' labelling	
Cosine similarity	0.7120
PCC similarity	0.6707
Jaccard similarity	0.6290
ICC similarity	0.6696
Labelling after applying accuracy measures compared to experts' labelling	
Average absolute deviation accuracy	0.6075
Normalized average absolute deviation	0.6075
Variance accuracy	0.6489
Variance by topic accuracy	0.6489
Standard deviation range accuracy	0.6489
Labelling after applying agreements measures compared to experts' labelling	
Modified agreement percentage	0.6079
Alpha agreement	0.6200
Labelling after applying majority consensus measures compared to experts' labelling	
Majority exact match	0.6079
majority class ratio	0.6079
Majority normalized class ratio	0.6079
Majority normalized class ratio algorithm	0.6079

Reflecting on the above table, we concluded that our model aligns more with the experts' ratings compared to the common majority voting method. The inferred labelling through similarity and accuracy measures outperformed the agreement values using simple majority voting, and reached a 0.71 agreement value using cosine similarity, which indicated "substantial" agreement with experts labelling, based on Landis & Koch's 1977 [74] agreement interpretation. The results presented in the table are based on the similarity and accuracy average rating methods. In general, it is found that applying average rating methods are superior to computing pairwise rating methods for both similarity and accuracy measures since pairwise rating did not aspire to better agreements or correlations compared to experts labelling. As using iterative algorithm with cosine similarity resulted in a slight change to labellers' weights, the labelling agreement values with experts remained the same.

As a result of these experiments, we propose computing the labellers' weights, utilizing all measurements and aggregating their weights. Accordingly, final dataset credibility labelling then is constructed based on labellers' aggregated weights. A comparison between derived tweet messages' labels $TLabel_{t_i}$using the labellers' aggregated weights method, which employ all the measurements proposed by this chapter, with experts rating values, is shown below in Table 4-22.

Table 4-22 Labelling after applying aggregation model compared to experts' labelling

Simple majority voting	**Expert1_A**	**Expert2_N**	**Expert3_E**	**Average**
Maj_Class2	0.6055	0.0418	0.21	0.2858
Maj_Low	0.5831	0.0814	0.1674	0.2773
Maj_Hi	0.6075	0.0294	0.2343	0.2904
Maj_N	0.6055	0.0418	0.21	0.2858
Aggregation model				
Agg_Model	0.6075	0.0294	0.2343	0.2904

After ranking the labellers with their weights, a correlation between their computed weights and the degree of their topic familiarity and interest has been calculated and the results showed a negative correlation (-0.022). In spite of this, we will rely exclusively on labellers' ratings for identifying the reliability as answers in the questionnaires cannot assure labellers' familiarity with the topic. Furthermore, it should be emphasized that after applying the proposed weights aggregation model, all the labellers who were identified with low propensity to trust were not from the top reliable labellers. In fact, three were identified from the moderate-reliability list and the others were retrieved from the low-reliability level list.

5 Credibility Detection Using Feature-based Approaches

This chapter elaborates on two important steps to detect credibility which are features extraction and credibility classification. In this study, we apply feature-based approaches to detect tweet messages credibility classes. With this approaches, the labelled set of messages where each message represented by a set of features is used by both, the statistical model and machine learning model to classify messages credibility. This detection approaches involves representing each labelled tweet message with measured features that are suitable for computing messages' credibility. We extracted and computed a wide range of features related to the messages' author and content; some of these features are novel while the majority has been proposed by previous studies in information credibility. Later on this chapter, statistical approach based on features frequencies, and machine learning classifier algorithm based on decision tree have been used to detect credibility of Arabic messages and identify credibility prominent features. The last section of this chapter reports on study of the effect of majority voting level on classification accuracy. It demonstrates that labelled dataset with higher level agreement between labellers can achieve better classification accuracy results.

5.1 Features Extraction and Evaluation

As features serve as good predictors of credible information, there is a wide range of features proposed in different studies to assess credibility of tweet messages. Most of these studies rely on Twitter features related to both: the messages' author and content [13]–[15], [17], [18], [38] for assessing information credibility. Yet some studies focus on the linguistic features of the content [8], [9], [12], [16] and others focus on a selection of features and check their usefulness and validity to predict credibility. A study was conducted by O'Donovan et al. 2012 [18], who examined how indicators such as retweet chain length and dyadic exchanges are used as metrics to measure credibility. Table 5-1 lists the used features by different studies along with their credibility prominent features.

Table 5-1 Existing credibility Twitter features

Used by	Features
Castillo et al. 2011 [13]	**Content, Author, Topic, and Propagation** **Prominent features: Topic:** fraction of tweets having URL, fraction of negative sentiment, fraction of tweets with !, and fraction of first-person pronoun. **Author:** friends count, statuses count, registration age, and followers count. **Propagation:** maximum level size of the retweet tree.
Gupta & Kumaraguru 2012 [14]	**Content and Author** **Prominent features: Content:** characters count, unique characters count, swear words count, inclusion of pronouns, presence of sad / happy emoticons, and presence of URL. **Author:** followers count, and username length.
Qazvinian et al. 2011 [16]	**Content and Author /Network** **Prominent features: Content:** log likelihood ratio (LL-ratio) content unigram, LL-ratio content bigram, LL-ratio content POS unigram, LL-ratio content POS bigram, LL-ratio URL unigram, LL-ratio ratio URL bigram, LL ratio hashtag. **All event users:** LL-ratio tweeting user, LL-ratio retweeted user. **Each feature is a log-likelihood ratio calculated against a positive (+) and negative (−) training models.**
Bhattacharya et al. 2012 [12]	**Prominent features: Content:** Unigram
Yang et al. 2012 [38]	**Content, Author, Propagation, Client, and Location** **Prominent features: Author**
Kang et al. 2013 [47]	**Content, Social/Author and Behavioural (Author + Friends)** **Prominent features for Dataset#1: Content:** news words, word count, presence of URL, sentiment positive and pronoun. **Author:** mutual friends count, ignoring friends count, status count, registration age, and listed count. **Behavioural (Author + Friends):** mean #hashtags, and mean #URLs in tweets. **Prominent features for Dataset#2:** **Content:** presence of URL, characters count, words count, news words, and number of mention. **Behavioural (Author + Friends):** mean # hashtags in tweets, # users that propagate the user, deviation of a user's retweet rate from the normal rate, # tweets propagated by other users, mean time between tweets, and mean response time to tweets.
Xia et al. 2012 [17]	**Content, Author, Topic, and Diffusion** **Prominent features: Content:** presence of URL, reply number, number of ?, number of !, number of @, length, words match keyword, is retweet. **User:** presence of description, followers count, friends count, verified, age, statues count. **Topic:** URLs fraction, hashtags fraction, address, address match topic address, number of positive/ negative.
Gupta et al. 2013 [37]	**Content and Author** **Prominent features: Content:** length, words count, contains ?, contains !, number of question marks, number of exclamation marks, contains happy/sad emoticon, contains first/second/third order pronoun, number of uppercase characters, number of negative/positive sentiment words, number of mentions, number of hashtags, number of URLs, retweet count
Kang, O'Donovan, & Höllerer 2012. [15]	**Content and Social/Author.** **Prominent features: Social/Author:** weighted combination of (deviation of a user's retweet rate from the average retweet rate, retweet rate deviation factored by number of followers and normalized by number of tweets, deviation of a user's followers from the mean number of followers normalized by number of tweets, deviation of a user's ratio of follower to following from the norm; the percentage of a user's tweets number on the topic to their total number of tweets) for a user on a given topic
Mendoza et al. 2010 [39]	**Content and Propagation** **Prominent features: Propagation:** retweet, **Content:** contain indicators of questioning
O'Donovan et al. 2012 [18]	**Content and Author/ Social/ Behavioural.** **Prominent features: Content:** URLs, mentions, retweets and tweet length

In this investigation, we extracted and computed a wide range of features related to the messages' author and content. The majority of features were proposed in previous studies in information credibility while some of them are unique. The new features include the use of dialect or religious words in the content, the use of names with news nature by the authors, the disclosure of users' information (such as education, occupation and contact details), and whether the authors' bios, names or locations are linked to the topic. As noted in Table 5-1, content features had a great impact on detecting credibility, hence this study concentrated more on employing content analysis techniques to compute and analyse the content data in both authors' data and the tweet messages' content. The following diverse techniques have been used to extract and compute some of the content features:

- **Arabic dialect words analysis:** To detect any dialect words in the tweet message content, a user-generated dictionary of colloquial Arabic "Mo3jam" has been used[13] as a lexicon for the informal words.
- **Arabic formal and cognition/saying words analysis:** For formal words, all cognition/saying verbs from a Linguistic Inquiry and Word Count (LIWC 2007) dictionary for Arabic texts dictionary [88] have been used in addition to phrases suggested by participants from the online survey.
- **Arabic pronouns analysis:** In this study, to detect the Arabic pronouns, the Linguistic Inquiry and Word Count (LIWC 2007) for Arabic texts dictionary [88] has been used (content_PronounsDNo), however all the available pronouns in the dictionary are only stand-alone-pronouns, whereas in Arabic, pronouns are frequently attached as suffixes to the nouns, verbs and particles. We can infer from a verb conjugation who the subject is, so it is not really necessary to use the subject pronoun in such cases except for emphasis. For this reason, another method was applied to detect pronouns by translating the messages to English and then calculating all singular 1st person (I/my/mine/me); 2nd person (you/your/yours); 3rd person (he/she/it/his/ her/hers/ its/him/). Plural 1st person (we/our/ours/us); 2nd person (you/ your/ yours); 3rd person (they/ their/ theirs/ them), which seems to be more effective (content_PronounsTNo).
- **Emotional analysis:** Subsequently the existing personal emotions' marks or words could be a good indicator of detecting online messages' credibility; hence we included a set of

[13] http://ar.mo3jam.com/dialect/Saudi

emotional features, including occurrences of emoticons, number of exclamation marks, number of question marks, and the occurrence of laughing words in the message content.

- **Durational analysis:** In addition, we extracted durational features, such as the length of tweet messages in words and in characters. Intuitively, people tend to believe that longer posts signify more information [89].

Based on the analysis of the credibility factors reported in previous Web credibility research, we found that there are some factors shared by most of these studies, and we can group these used features into four main factors:

- authority and topical expertise (of the source)
- data quality (of the content),
- and popularity (of the content and the source).

A total of 46 features has been extracted and computed in this study, comprised of 24 content features and 22 author features. Used features with their related credibility factors have been listed below in Table 5-2. It should be noted that selected features were grouped and treated as new features such as (author_AllRelate: if author's location/ bio/ name related to the topic, author_AllInf: if author's bio has education/work/ contact details).

Table 5-2 Used credibility Twitter features

Features	Type	Factor
1. topic: content topic genera	content	data quality
2. content_Rank: tweet message date	content	data quality
3. content_RetweetNo: number of retweet	content	popularity
4. content_FavNo: number of favorites	content	data quality
5. content_HashNo: number of hashtags	content	popularity
6. content_SpellNo: number of spelling errors	content	data quality
7. content_QmarkNo: number of question marks	content	data quality
8. content_ExcmarkNo: number of exclamation marks	content	data quality
9. content_EmotiNo: number of emoticons	content	data quality
10. content_SpecialchNo: number of special characters	content	data quality
11. content_CharNo: number of characters	content	data quality
12. content_WordsNo: number of words	content	data quality
13. content_HasURL: if content has a reference	content	data quality
14. content_HasImage: if content has image attached	content	data quality
15. content_PronnounsTNo: number of pronounces - English translation	content	data quality
16. content_PronounsDNo: number of pronouns - LIWC2007 dictionary	content	data quality
17. content_SQuest: if content start with questions	content	data quality

18. content_HasLaugh: if content has laughing words (loool, هههه, لوول)	content	data quality
19. content_ DialectWNo: number of dialects words - Mo3jam dictionary	content	data quality
20. content_BadSwearNo: number of bad/swear words - Mo3jam dictionary	content	data quality
21. content_ReligiousWNo: number of religious words - LIWC2007 dictionary and Mo3jam dictionary	content	data quality
22. content_AllDial: if content has dialects/ bad/ swear / religious words (group feature)	content	data quality
23. content_Formal: number of formal words - cognition/saying verbs -LIWC2007 dictionary	content	data quality
24. content_HasUrgnews: if content has urgent words (urgent, عاجل)	content	data quality
25. author_Verif: if author has verified account	author	authority
26. author_DefImage: if author used Twitter default image (Egg)	author	authority
27. author_FwngNo: author's following number	author	popularity and authority
28. author_FlrNo: author's followers number	author	popularity and authority
29. author_LogFlrNo: logarithm of the followers number	author	popularity and authority
30. author_RatioFwFl: following – follower ratio	author	popularity and authority
31. author_TweetsNo: author's tweets number	author	authority and topical expertise
32. author_FavNo: author's favorites number	author	authority and topical expertise
33. author_RatioTweetFav: tweets – favorites ratio	author	authority and topical expertise
34. author_News: if author's has a name or profile bio with news nature	author	authority and topical expertise
35. author_HasBio: if author has a bio	author	authority and topical expertise
36. author_Edu: if author's bio has education inf.	author	authority and topical expertise
37. author_Emp: if author's bio has position of employment	author	authority and topical expertise
38. author_Contact: if author's bio has contact details	author	authority
39. author_AllInf: if author's bio has education/work/ contact details (group feature)	author	authority and topical expertise
40. author_AllInf2: if author's bio has at least two of these inf.: education/work/ contact details (group feature)	author	authority and topical expertise
41. author_HasWeb: if author has a webpage	author	authority
42. author_YearsNo: years since joined Twitter	author	authority
43. author_DescRelate: if author's bio and name related to the topic	author	authority and topical expertise
44. author_LocationRelate: if author location related to the topic	author	authority and topical expertise
45. author_AllRelate: if author's location/ bio/ name related to the topic (group feature)	author	authority and topical expertise
46. author_HasSpecialch: if author's bio content has any special characters	author	authority

Below in Table 5-3 is an example of extracted and computed features values for a sample tweet message:

"الاشتباه بإصابة ٣ أطباء وممرضة بـ " #كورونا " في #مستشفى_الملك_فهد بـ #جدة"

http://t.co/U7HYosGCna "

from the Topic#1: Crises - Health - Domestic - Corona virus in Saudi Arabia (فيروس كورونا في السعودية) April 2014.

Table 5-3 Extracted and computed features' values for a sample tweet

Features	Features' values	Features	Features' values
topic_Genera	1: Crises – Health - Domestic - April 2014 / فيروس كورونا في السعودية	**tweeted_by_Author**	dr3lo
tweet_Content	"الاشتباه بإصابة ٣ أطباء وممرضة بـ " #كورونا " في #مستشفى_الملك_فهد بـ #جدة http://t.co/U7HYosGCna	**Author_Image_File**	http://pbs.twimg.com/profile_images/425732799199735808/ZqxF-gxr_normal.jpeg
tweet_Date	8 April 2014	**author_Description**	internship doctor طبيب امتياز #scorpion #ALAHLI, coffee-drinker ,Instgram :dr_3loo kik:dr.3lii
content_URL	http://twasul.info/48634/	**author_Location**	jeddah
content_Domain	twasul.info	**author_Web**	NA
content_Rank	5	**author_TimeZone**	Hawaii
content_RetweetNo	10	**author_Joined**	23 April 2012
content_FavNo	0	**author_Verif**	0
content_Hashtags	كورونا مستشفى_الملك_فهد جدة	**author_DefImage**	0
content_HashNo	3	**author_FwngNo**	220
content_SpellNo	0	**author_FlrNo**	745
content_QmarkNo	0	**author_LogFlrNo**	2.8722
content_ExcmarkNo	0	**author_RatioFwFl**	0.3
content_EmotiNo	0	**author_TweetsNo**	3154
content_SpecialchNo	0	**author_FavNo**	195
content_CharNo	98	**author_RatioTweetFav**	16.17435897
content_WordsNo	14	**author_News**	0
content_HasURL	1	**author_HasBio**	1
content_HasImage	0	**author_Edu**	0
content_PronounsTNo	0	**author_Emp**	1
content_PronounsDNo	0	**author_Contact**	0
content_SQuest	0	**author_AllInf1**	1
content_HasLaugh	0	**author_AllInf2**	0
content_DialWNo	0	**author_HasWeb**	0
content_DadSwearNo	0	**author_YearsNo**	2
content_ReligiousWNo	0	**author_DescRelate**	1
content_AllDeli	0	**author_LocationRelate**	1
content_Formal	0	**author_AllRelate**	1
content_HasUrgnews	0	**author_HasSpecialch**	0

5.1.1 Arabic language characteristics

Arabic language has special characteristics that should be taken into consideration while making text analysis. The following listing outlines some Arabic language characteristics which might hinder processing and analysing Arabic content [90]–[94]:

- Arabic has several diacritics (vowels); as opposed to English, these diacritics are rarely used in writing. **(Difficulty with multiple semantics)**
- Arabic is a highly inflected language and very rich in both vocabulary and morphological variation. **(Less word occurrence - NLP)**
- Arabic language does not use capital letters as in English. **(Hardens to extract proper nouns - Named Entity Recognition, capital letters could be also used with emotional features as indication for an emphasis)**
- Arabic words are shorter than English words. **(Word-length features are less discriminating)**
- Arabic words are sometimes elongated for purely stylistic reasons, using a special character that resembles a dash (--). **(Elongation can affect word-length feature)**

In evaluating the English and Arabic content using machine learning techniques and feature analysis, research from different fields found key differences between the language models. Some interesting observations are presented in Table 5-4 below from studies that include sentiment analysis, spam detection and authorship identification. All studies indicate that selecting suitable features for classification problems depends primarily on the language of the examined content. Moreover, some content features that are studied in previous work for English posts, such as the amount of character capitalization does not apply to the Arabic language. Therefore it is recommended to identify previously studied English content features and check if it could be substituted with other features related to Arabic content.

Table 5-4 Arabic and English language models

Used by	Arabic and English Differences
Abbasi et al. 2008 [91] **Sentiment Analysis - English and Arabic Web forum postings.**	▪ Effective features for Arabic: (short words, total char, elongation , long words, digits, vocabulary richness) where for English messages: (total char., $, &, {, digit count, words “therefore”, word “however”, word “nevertheless”) ▪ There is a significant usage of fonts, colors, elongation, numbers, and punctuation features in Arabic forums. ▪ Number of n-grams used for the English feature set is nearly threefold those used for the Arabic. ▪ In general, English features had higher usage rates than the Arabic feature set.

Abbasi & Chen 2005 [92] **Dark Web Authorship identification - English and Arabic Web forum postings.**	▪ (Word length) feature was more effective in English messages as compared to Arabic messages. ▪ Arabic messages tended to be considerably longer than the English messages and had a more formal structure, featuring more greetings, more sentences, and lengthier paragraphs. ▪ Arabic messages used a plethora of font colors and sizes; in contrast to the English messages, where fonts featuring black, 10-to-12- point size. ▪ Arabic messages had higher frequency of embedded images than the English messages (approximately 20 times more). ▪ Arabic messages had many more links to static, dynamic, and image pages. ▪ Author contact information was seldom provided in Arabic.
Alarifi & Alsaleh 2012 [90] **Web Spam - Webpages**	▪ In Arabic dataset, (amount of anchor text in the Web page and number of images in the Web page) features gave the best performance, where (number of words in the <meta> element and number of words in the Web page) features gave the best detection performance in the English dataset.
Al-Kabi, et al. 2012 [93] **Web Spam - Webpages**	▪ The weights of (number of characters/words in the <title> element; number of characters/words in the <meta> element; number of popular words in the pages; and number of characters/words, in the <body> element) features are higher than their counterparts' weights for English spam web pages.
Goweder & De Roeck 2001 [94] **NLP - Word Occurrence**	▪ English words are repeated more often than Arabic ones for the same text length.

5.2 Credibility Assessment Using Statistical Approach

In previous studies, most features were proposed to assess credibility of English tweet messages and consequently their practicality might vary within the Arabic context. To understand the strength of the detection features, we will analyse features distribution in our dataset and attempt to identify the best features that detect credibility. After applying the frequency of features model used by O'Donovan et al. 2012 [18], we use results from both, statistical applied model and survey study, to identify the features that have more influence on credibility perception. In the next sections, we elaborate on methods used to identify the prominent features: 1) using feature frequencies around their average across different credibility classes and 2) by investigating survey results of how important author and content features to participants using a Likert response scale between 1 and 5 [95].

5.2.1 Evaluating features using relative frequency

Inspired by work done by O'Donovan et al. 2012 [18], we used a histogram as shown in Figure 5-1 to illustrate the distribution of features across three classes {1,2,3}. First we partitioned the range of computed feature values around their average - meaning, dividing the

entire range of features' values into two intervals - and then calculating relative frequency to measure the proportion or percent of the data values into the two intervals for each credibility class.

$$Relative\_Frequency = \frac{Features' Frequency}{Sample\ size}$$

Moreover, histogram is used to display the relative frequencies of features using resulting labelling from proposed labellers' ranking method. It demonstrates the proportion of instances that fall into the two intervals, with the total area equalling 1. Detailed table shown the relative frequencies of features across three classes {1, 2, 3} is shown in Appendix C, Table C-1

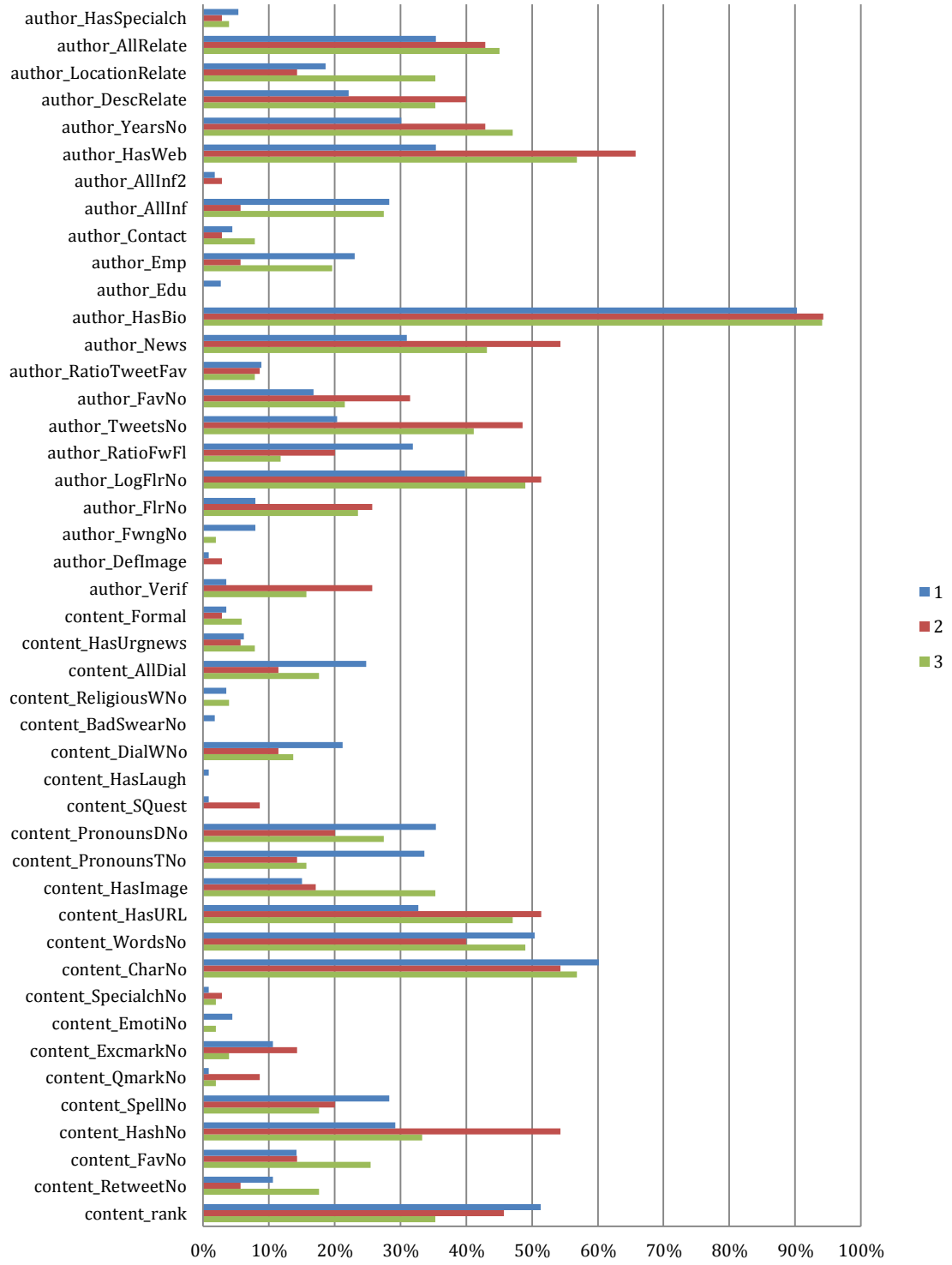


Figure 5-1 The distribution of features across three classes {1, 2, 3}

Several general conclusions can be drawn from the abovementioned chart in Figure 5-1. Firstly, the usage of features appears greater in the second questionable class {2}, which make it difficult to distinguish a major difference for features distribution between both low-credibility {1} and high-credibility {3} classes. Generally, the main features that distinguish the high-credibility class from other classes are the use of Twitter default image (Egg) by the authors and

if message content starts with a question. Furthermore, adding image to the content is clearly present in class {3} where the use of exclamation marks appears more often on {1, 2} classes.

The chart also shows that messages' posting date and time influences credibility perception; in other words, more current posts result in greater credibility. Accordingly, features such as higher number of retweets and favourites, more formal cognition/saying words, less ratio value to following-followers numbers, old Twitter account, and if author location in relation to the topic; all positively affect messages' credibility. To provide a clear illustration for features occurrences on high-credibility {3} and low-credibility classes {1}, the features distribution between only these classes is presented in Figure 5-2.

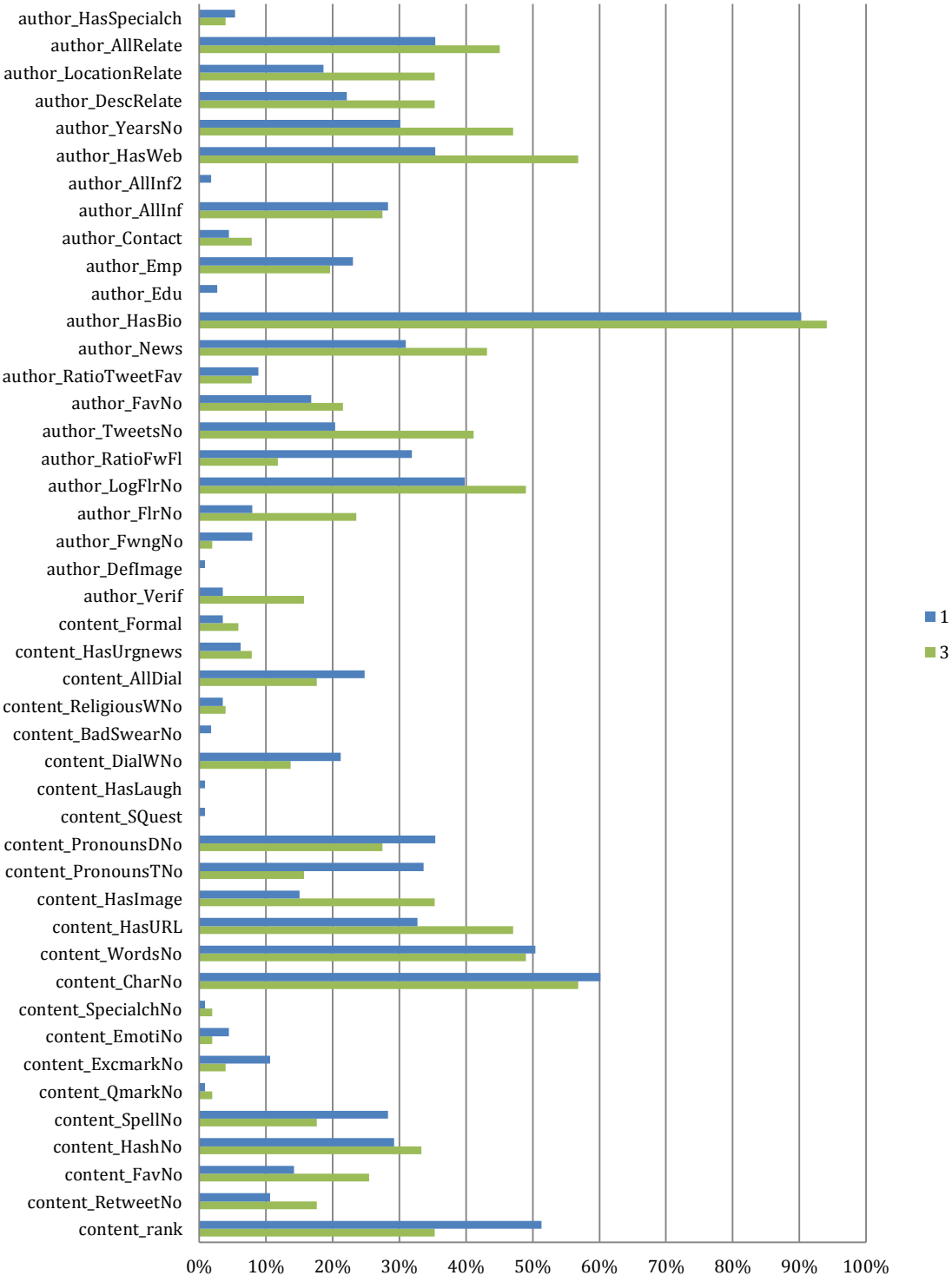


Figure 5-2 The distribution of features across two classes {1, 3}

From Figure 5-2 above, we observed that presenting messages with low data quality content resulted in a low credibility perception. In fact, features with higher representation in low-credibility class are frequently related to data quality factor that covers the textual content and the writing style. Features such as the availability of swearing/ inappropriate and humorous words in the content, and the use of more informal dialectal words might be indicators for low-

credibility messages. Similarly, features such as the number of spelling errors, exclamation marks, emoticons, and pronouns might be used to distinguish low-credibility messages from other messages. In addition, features related to the authority of the source like: the use of Twitter profile default image, following large number of users where the ratio between the following and followers is high are all possible signals for low-credibility messages.

On the other hand, for higher credibility messages, features which are related to the authority of the source were utilized more by this class. For instance, features such as having "verified" account, large number of followers, favourites and tweets, having a webpage, having an old Twitter account, having a bio related to the discussed topic, all appeared to be signs for higher credibility. While it is worth mentioning that adding a reference or image to the content makes it more plausible, it appears that attached images are considered stronger evidence for higher credibility. In comparison with other studies, similar features such as large number of retweets, favourites, and hashtags number are also signs for higher credibility messages in this study.

Surprisingly, findings regarding authors' bio information, where the specific chart suggested that assuming a higher degree of authors' self-disclosure - such as stating education level and employment, didn't not positively influence the trustworthiness of messages. Twitter users' authority and credibility may be judged by other factors, rather than their personal information provided on their bios. However, adding Webpage and contact information might offer a degree of authority. In regard to the length of tweet messages in words and characters features, the chart suggests neither number of words nor number of characters features is capable to distinguish the credibility level; as stated previously, word-length features are less discriminating as Arabic words are shorter than English words.

5.2.2 Evaluating features using survey results

A phase from the online study advises participants to rate the importance of different features on assessing the information credibility of Twitter messages. The majority of the presented features are extracted from a previous study [53] . Results from this phase are reported as an ordered list (see Table 5-5) of the most prominent features that indicate the influence of the authors and the quality of the tweets.

Table 5-5 Prominent features – survey results

Features	Average
Author has a "verified" account	4.0357
Author used his/her real name	3.9048
Author used organizational name	3.7727
Author used professional name	3.6364
Content with formal language (No spelling or grammar mistakes)	3.6071
Author often tweets on specific topic	3.5455
Content with URL	3.5185
Content with event "image" attached	3.5
Author is known (personal/someone you've heard of)	3.5
Author location near news event topic	3.4545
Author used his/her real photo	3.4286
Author has many followers	3.4286
Author Twitter bio include contact information	3.381
Author Twitter bio include position of employment	3.3636
Author Twitter bio include Organizational authorship	3.3636
Author Twitter bio suggests topic expertise	3.3182
Author has added "WebPage"	3.2963
Author is followed by you	3.2857
Author is known (celebrity)	3.2273
Content is similar with many tweets	3.1818
Author Twitter bio include Education level	3.1818
Content with more "favorite"	3.1071
Author used topical name	3.0909
Author used image represent organization	3.0909
Author often mentioned/retweeted	3.0909
Content with more retweets	3.0769
Content posted recently	2.9090
Content with hashtags	2.8929
Content with more mentions "@"	2.8889
Author used image represent profession	2.8182
Author has an old Twitter account	2.7895
Author is following many users	2.7727
Content with longer length – more characters and words	2.75
Content with exclamation mark "!"	2.6429
Content with question mark "?"	2.6071
Content with personal pronouns	2.3929
Content with emoticons	2.0714
Author used any image	2.0455
Author used nick name/ any name	1.8636
Content with unique special characters	1.8571
Author has Twitter profile image (Egg)/no image	1.8182
Content with swear/bad words	1.4286

Based on the preceding table above, features related to the authority and expertise of authors such as: having a "verified" account, using identifiable name either a real name,

organizational name, or professional name, always tweeting on specific topics, being known, using real photo, and having many followers, all were rated as important features that indicate higher credibility. In addition, features related to acquiring better data quality of content such as: using formal language with no spelling or grammatical errors and using supplementary information (such as references and images), also indicate higher credibility for participants. Simultaneously, features associated to having low quality of content such as using swear words, emoticons, pronouns, question marks and exclamation marks have been rated as low-credibility signals as it appeared at the end of the list. Yet again, low-source authority features such as using nicknames, assigning the Twitter profile image or any random image for the source bio, and following many users are indicators for low credibility.

Strangely enough, participants consider having a long content with more words and characters as a low-credibility sign although it is a basic indicator of an informed message. Another observation involved authors' bio information, where the participants suggested that including authors' work and education details would be a signal for higher credibility, however that was not supported with the results from the statistical model.

5.2.3 Labellers' similarity and agreement compared to messages' features occurrences

This section investigates the assumption whether the most similar and agreed labellers share similar credibility features. We identified the most related labellers in different similarity and agreement measures and examined the feature distributions using their assigned labelling. The highest pairwise rating similarity and agreement values between labellers for the following measures: pearson similarity, cosine similarity, jaccard similarity, and alpha agreement are presented in Appendix C, Table C-2.

- **Using Pearson similarity:** By applying pearson similarity for every pair of labellers, we identified that labeller#12 and labeller#16 illustrate the highest pearson correlation with a value reached: 0.805. Figure 5-3 shows the features distribution for tweet messages within class {2} only as labeller#12 did not assign any tweet with credibility class {3}, also labeller#16 did not assign any tweet with credibility class {1}. A detailed result in Appendix C, Table C-3 shows the messages' features occurrence between the most similar labellers using pearson similarity.

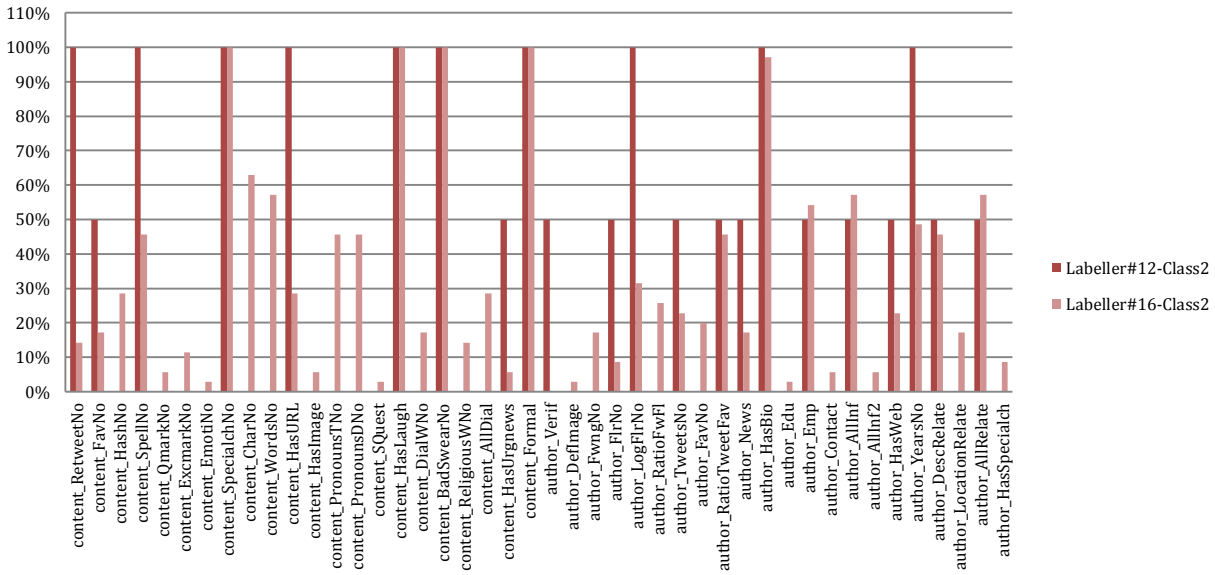
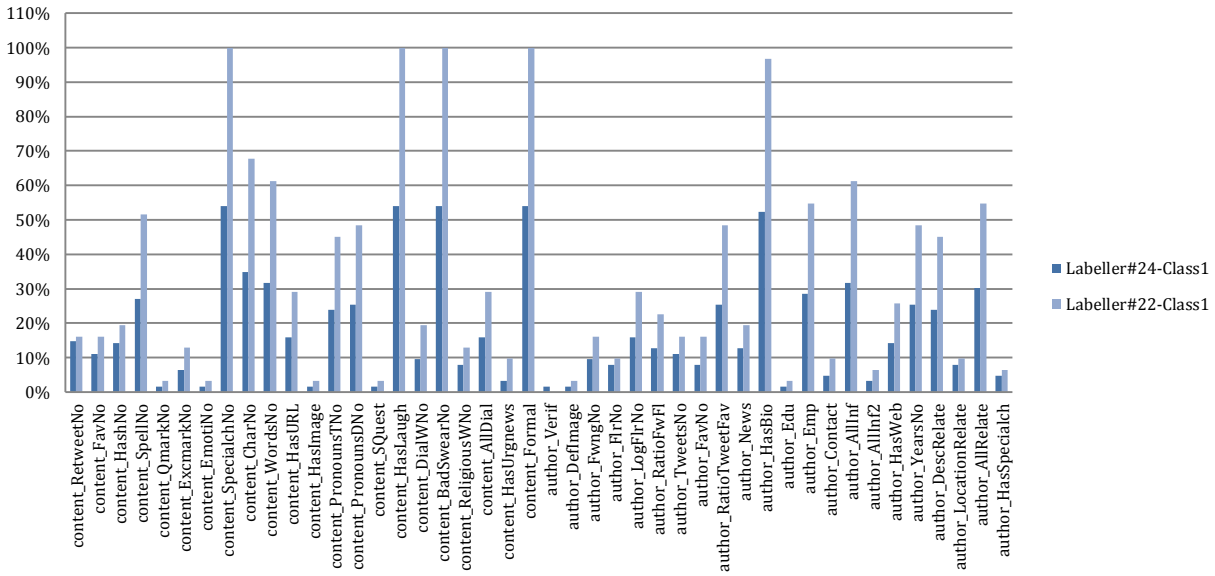


Figure 5-3 The distribution of features across labeller#12 and labeller#16

- **Using Cosine and Jaccard similarity:** For both cosine and jaccard similarity measures, labeller#22 and labeller#24 were identified with the highest similarity values. For cosine similarity, their computed value was: 0.9562, and for jaccard similarity, it was: 0.897. Features distribution for the shared tweet messages between both labellers are presented in Figure 5-4. For the detailed results, Table C-4 in Appendix C shows the tweet messages' features distribution percentages across the most similar labellers using both cosine and jaccard similarity.



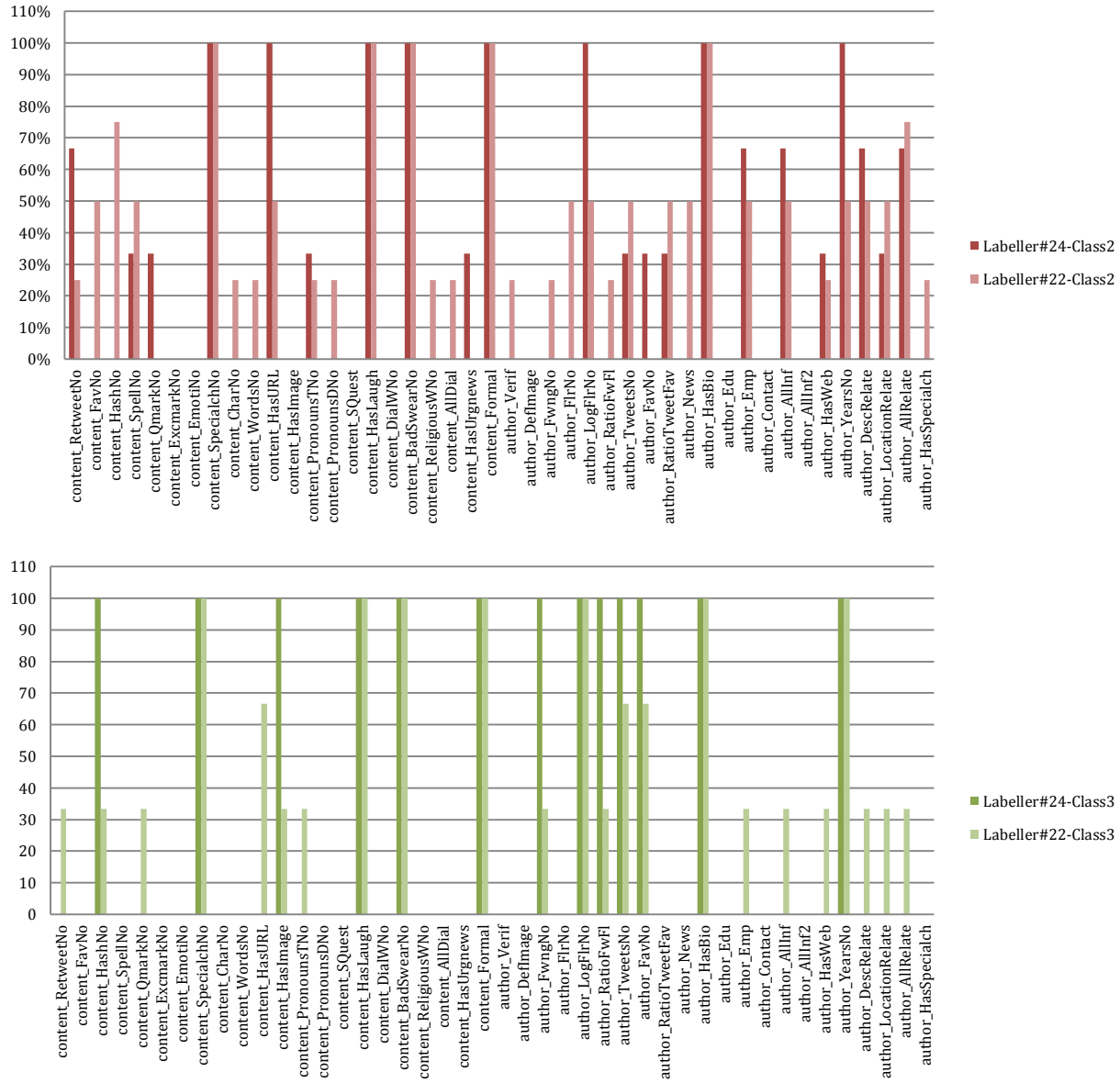
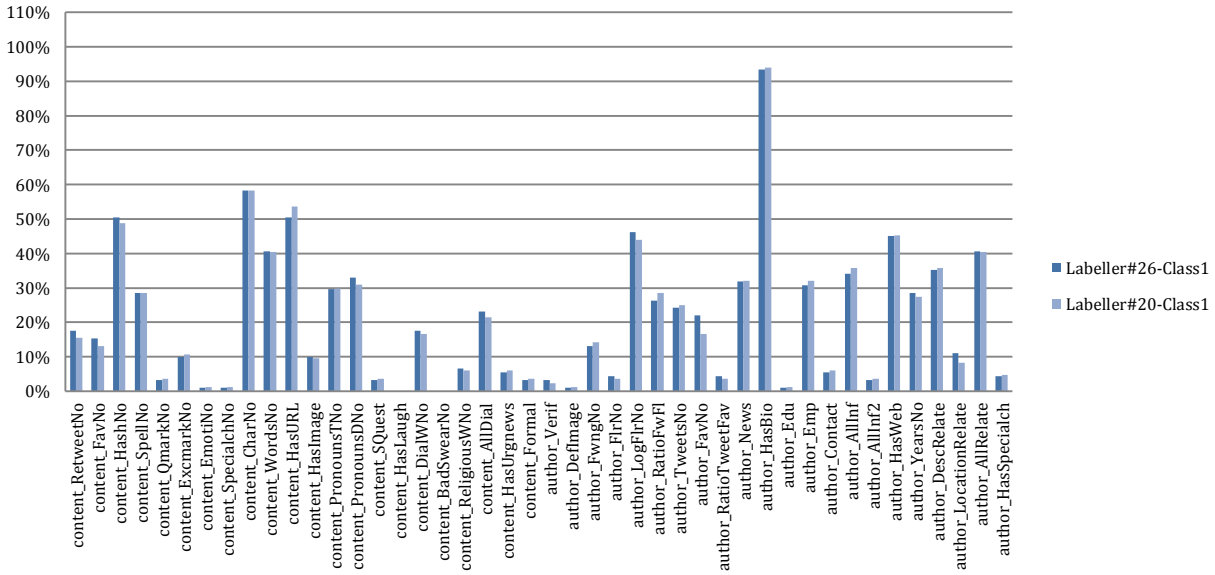
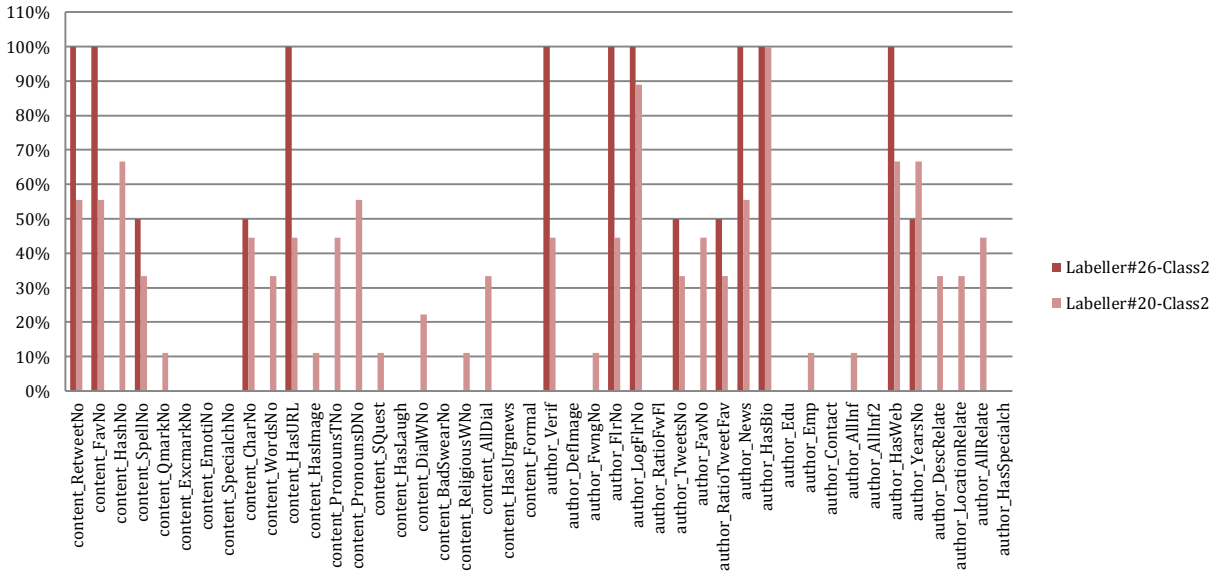


Figure 5-4 The distribution of features across labeller#22 and labeller#24

- **Using Alpha agreement:** We concluded that labeller#20 and labeller#26 have the highest agreement value among other labellers; their agreement value was: 0.66. Features distribution for shared tweet messages within classes {1,2,3} are presented in Figure 5-5, where the detailed features distribution data is presented in Appendix C, Table C-5.


110%
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
content_RetweetNo
content_FavNo
content_HashNo
content_SpellNo
content_QmarkNo
content_ExcmarkNo
content_EmotiNo
content_SpecialchNo
content_CharNo
content_WordsNo
content_HasURL
content_HasImage
content_PronounsTNo
content_PronounsDNo
content_SQuest
content_HasLaugh
content_DialWNo
content_BadSwearNo
content_ReligiousWNo
content_AllDial
content_HasUrgnews
content_Formal
author_Verif
author_DefImage
author_FwngNo
author_FlrNo
author_LogFlrNo
author_RatioFwFl
author_TweetsNo
author_FavNo
author_RatioTweetFav
author_News
author_HasBio
author_Edu
author_Emp
author_Contact
author_AllInf
author_AllInf2
author_HasWeb
author_YearsNo
author_DescRelate
author_LocationRelate
author_AllRelate
author_HasSpecialch
Labeller#26-Class1
Labeller#20-Class1


110%
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
content_RetweetNo
content_FavNo
content_HashNo
content_SpellNo
content_QmarkNo
content_ExcmarkNo
content_EmotiNo
content_SpecialchNo
content_CharNo
content_WordsNo
content_HasURL
content_HasImage
content_PronounsTNo
content_PronounsDNo
content_SQuest
content_HasLaugh
content_DialWNo
content_BadSwearNo
content_ReligiousWNo
content_AllDial
content_HasUrgnews
content_Formal
author_Verif
author_DefImage
author_FwngNo
author_FlrNo
author_LogFlrNo
author_RatioFwFl
author_TweetsNo
author_FavNo
author_RatioTweetFav
author_News
author_HasBio
author_Edu
author_Emp
author_Contact
author_AllInf
author_AllInf2
author_HasWeb
author_YearsNo
author_DescRelate
author_LocationRelate
author_AllRelate
author_HasSpecialch
Labeller#26-Class2
Labeller#20-Class2

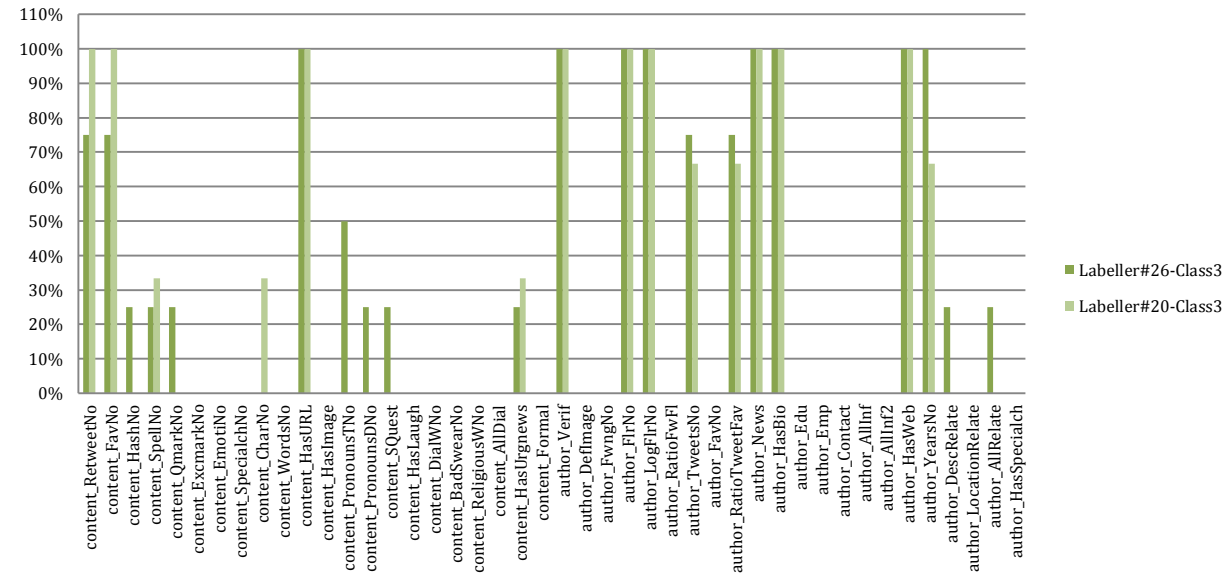


Figure 5-5 The distribution of features across labeller#20 and labeller#26

A number of observations can be detected from all of the charts directly above: Generally, having similar labellers does not mean assigning credibility judgments using the same features. Shared features between labellers mainly appear with the most agreed labellers using Krippendorff's alpha agreement whereby almost 100% features similarity for class {1}. For all used similarity and agreement measures, features in common between labellers are largely present in classes {1, 3}, but moderately in the questionable class {2}. It is reasonable that the features differences between labellers appear more in class {2} as it the questionable class where labellers couldn't reach to decisive credibility judgement. Table 5-6 lists all the shared features between the most similar and agreed labellers along with an estimate of the features similarity percentage.

Table 5-6 Shared features between the most similar and agreed labellers

Shared features	**Class {1}**	**Class {2}**	**Class {3}**
Using pearson similarity	NA	content_SpecialchNo, content_HasLaugh, content_BadSwearNo, content_Formal, author_HasBio **≈ 11%**	NA
Using cosine and jaccard similarity	content_RetweetNo, content_QmarkNo, content_EmotiNo, content_HasImage, content_SQuest, author_Verif, author_DefImage, author_FlrNo,	content_ExcmarkNo, content_EmotiNo, content_SpecialchNo, content_HasImage, content_SQuest, content_HasLaugh, content_DialWNo, content_BadSwearNo,	content_FavNo, content_SpellNo, content_ExcmarkNo, content_EmotiNo, content_SpecialchNo, content_CharNo, content_WordsNo, content_PronounsDNo, content_SQuest, content_HasLaugh, content_DialWNo, content_BadSwearNo, content_ReligiousWNo, content_AllDial, content_HasUrgnews, content_Formal,

	author_Edu, author_LocationRelat, author_HasSpecialch **≈ 25 %**	content_Formal, author_DefImage, author_HasBio, author_Edu, author_Contact, author_AllInf2 **≈ 31 %**	author_Verif, author_DefImage, author_FlrNo, author_LogFlrNo, author_RatioTweetFav, author_News, author_HasBio, author_Edu, author_Contact, author_AllInf2, author_YearsNo, author_HasSpecialch **≈ 63 %**
Using alpha agreement	All features **≈ 100 %**	content_ExcmarkNo, content_EmotiNo, content_SpecialchNo, content_HasLaugh, content_HasUrgnews, content_BadSwearNo, content_Formal, author_DefImage, author_RatioFwFl, author_Edu, author_Contact, author_AllInf2, author_HasSpecialch **≈ 29 %**	content_ExcmarkNo, content_EmotiNo,content_SpecialchNo, content_WordsNo, content_HasImage, content_HasLaugh, content_DialWNo, content_BadSwearNo, content_ReligiousWNo, content_AllDial, content_Formal, author_Verif, author_DefImage, author_FwngNo, author_FlrNo, author_LogFlrNo, author_RatioFwFl, author_News, author_HasBio, author_Edu, author_Emp, author_Contact, author_AllInf, author_AllInf2, author_HasWeb, author_LocationRelate, author_HasSpecialch **≈ 61 %**

5.3 Credibility Assessments Using Machine Learning Approach

The final dataset consisting of the labelled set of messages and each message represented by a set of features is used by the classifier to train its model and acquire necessary knowledge which will be used to classify the messages credibility. We built a three way classifier-based feature for classifying the credibility into three classes {1,2,3} and the machine learning tool Weka [49], developed and maintained by the University of Waikato -New Zealand, has been used for this purpose. Weka is an open source java based machine–learning workbench that offers a large number of machine learning classification algorithms with additional tools for performing pre-processing tasks. In this study, we used decision tree classifier algorithm J48 (Open Source Java implementation of the C4.5 decision tree algorithm [48] in Weka data mining tool [49]) to carry out the credibility classification as it is a widely known algorithm that has been shared by previous credibility classification studies. Decision tree algorithm is a predictive model structured as a tree where leaf nodes represent class labels and branches represent features that lead to right class labels.

5.3.1 Building classification model

In this section, the study presented the classification accuracy results after applying decision tree algorithm J48. To prepare the dataset for classification, a Weka dataset using Attribute-Relation File Format (ARFF) was created for this experiment and presented in Table 5-7. In

Weka, we chose J48 as our classifier algorithm, decision tree is the main technique used in the J48 algorithm.

Table 5-7 Classification Weka data

```
@attribute topic numeric
@attribute content_rank numeric
@attribute content_retweetno numeric
@attribute content_favno numeric
@attribute content_hashno numeric
@attribute content_spellno numeric
....

@attribute author_verif numeric
@attribute author_defimage numeric
@attribute author_fwngno numeric
@attribute author_flrno numeric
@attribute author_logflrno numeric
@attribute author_ratiofwfl numeric
.....

@attribute Agg_Model {1,2,3}

@data

1,1,146,13,2,1,0,0,0,0,106,11,1,0,0,0,0,0,0,0,0,0,0,0,1,0,29,366215,5.5637,0,8467,0,0,1,1,0,0,0,0,0,1,2,0,0,0,0,3,3
1,2,1,0,1,0,0,0,0,0,140,21,0,0,0,1,0,0,0,0,2,1,1,0,0,0,287,75,1.8751,3.83,1265,122,10.36885246,0,1,1,0,0,1,0,0,2,0,1,1,0,1,1
1,3,1,0,1,0,0,0,0,0,67,6,1,0,0,0,0,0,0,0,0,0,0,0,0,7,12,1.0792,0.58,294,3,98,0,0,0,0,0,0,0,0,1,0,0,0,0,1,1
1,4,13,0,2,1,0,0,0,0,109,11,1,0,0,0,0,0,0,0,0,0,1,0,0,0,45,4674,3.6697,0.01,1037,2,518.5,0,1,0,1,0,1,0,0,2,1,0,1,0,1,1
1,5,10,0,3,0,0,0,0,0,98,14,1,0,0,0,0,0,0
.....
```

In order to have a fair measure of the performance of the classifier; we used a cross-validation with 10 folds. K-cross-validation procedure is the most common validation procedure used in information retrieval domain for evaluating the performance of a classifier. In k-cross-validation, the data is randomly divided into k equal, or close to equal subsets, usually 10. Each subset is predicted via the classification rule constructed from the remaining (k-1) subsets. This process is repeated until we have used every subset as the testing subset (Figure 5-6). To measure the performance, the error rate (which is the number of incorrectly predicted instances divided by the total number of test instances) was evaluated for every k fold. Then, the k error rates were averaged to obtain the final result.

$$Error = \frac{1}{k}\sum_{i=1}^{k} Error_i$$

The advantage of this procedure is that the error rate is an unbiased estimate as it makes use of all data instances for both training and testing, purportedly to derive a more accurate and consistent estimate of model classification performance. Although, k value is unfixed parameter and different values for k can be used, in practice, 10 have been found a reasonable estimate of error rate [96].

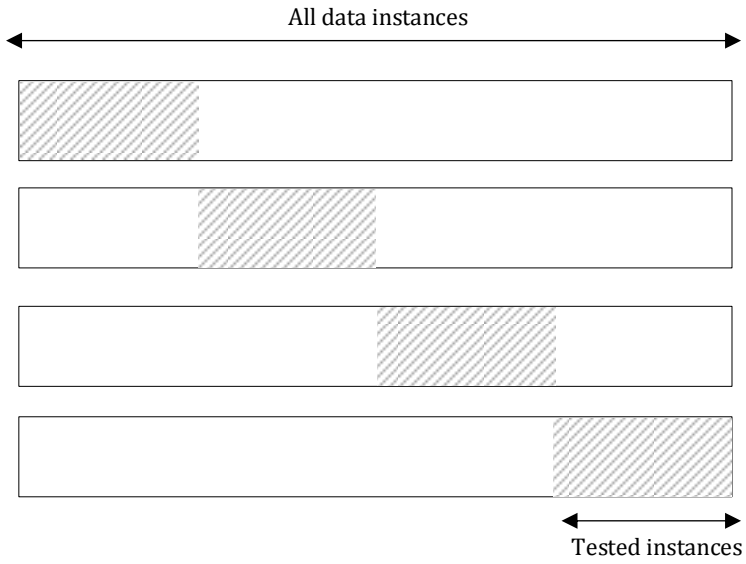


Figure 5-6 K-cross-validation

5.3.1.1 Decision tree classifier

The classification decision tree literally creates a tree with branches, nodes, and leaves. Decision tree is constructed in a top-down recursive divide-and-conquer approach. So, given a tweet message for which the associated class label is unknown, the feature values of the tweet message are tested against the decision tree. A path is traced from the root to a leaf node and a class label is assigned. Basic notions for the widely known classifier decision tree [97] are presented below in Table 5-8.

Table 5-8 Basic notions for the classifier decision tree

Decision Tree C4.5: Generate a decision tree
Input:
D: a set of training tuples and their associated class labels.
attribute_list: a set of candidate attributes.
Attribute_selection: a procedure to determine the *splitting_criterion* that "best" partitions the tuples into individual classes. This criterion consists of a *splitting_attribute* and, possibly, either a *split point* or *splitting subset*.
Output: A Decision Tree
1) create a node *N*;
2) if tuples in *D* are all of the same class, *C* then

3) return *N* as a leaf node labelled with the class *C;*

4) if *attribute_list* is empty then

5) return *N* as a leaf node labelled with the majority class in *D;*

6) apply *Attribute_selection* method (*D, attribute_list*) to find "best" *splitting_criterion*;

p_i = *|Ci,D|/|D| // the probability that arbitrary tuple in D belongs to class Ci*

Expected information (entropy) needed to classify a tuple in D: $Info(D) = -\sum_{i=0}^{m} p_i \log_2(p_i)$

Information needed to classify D after using A to split D into v partitions: $Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$ *// if A is discrete-valued.*

$\frac{a_i + a_{i+1}}{2}$*, If A is continuous-valued,*

Information gained by branching on attribute A: Gain (A) $= Info(D) - Info_A(D)$

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

7) label node *N* with *splitting_criterion*;

8) if *splitting_attribute* is discrete-valued and multiway splits allowed then // not restricted to binary trees

9) *attribute_list* ← *attribute_list* – *splitting_attribute*; // remove splitting attribute

10) for each outcome *j* of *splitting_criterion* // partition the tuples and grow subtrees for each partition

11) let D_j be the set of data tuples in *D* satisfying outcome *j*; // a partition

12) if D_j is empty then

13) attach a leaf labelled with the majority class in *D* to node *N*;

14) else attach the node returned by Generate decision tree (D_j, *attribute_list*) to node *N*;

end for

15) return *N*;

Other classification algorithms may also apply, the data mining tool Weka offers a large number of machine learning classification algorithms where we could use some in this research study.

5.3.1.2 Classifier accuracy

Assessing the performance of classification is an important aspect to determine the effectiveness of used model. Different performance measures are presented here as the metrics to identify the usefulness of the credibility model.

- **Accuracy rate:** The accuracy rate is the standard measure reported for the assessment of classification performance. It is the proportion of correctly classified instances to the total number of test instances. Usually it is presented as confusion matrix where the cells at the diagonal represent the classes with correct predictions, while other cells show misclassifications.
- **Kappa:** is a chance-corrected measure of agreement between the classifications and the true classes. It is calculated by subtracting the agreement expected by chance from the

observed agreement and dividing by the maximum possible agreement. A value greater than 0 means that classifier is doing better than chance.

- **The precision and recall:** Precision can be seen as a measure of exactness, whereas recall is a measure of completeness. Given a class, recall is the proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate), whereas precision is the proportion of instances that are truly of a class divided by the total instances classified as that class.

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

True positive (TP) represents correct classifications (instances correctly classified as a given class), while false positive (FP) represents misclassifications (instances falsely classified as a given class). The F1 is a combined measure for precision and recall and can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0.

$$F1 = 2 . \frac{Precision . Recall}{Precision + Recall}$$

5.3.2 Classification results

In this section, we listed all the results for classification mainly using accuracy rate measure. We evaluated three types of datasets: 1) labelled dataset obtained using simple majority voting, 2) dataset with labels values obtained after applying selective labellers' weighting measures, and 3) labelled dataset obtained with proposed labellers' weighting aggregation model.

5.3.2.1 Classification results before applying the proposed model

Messages' credibility classification results for labelled dataset obtained using simple majority voting options: Maj_Class2, Maj_Low, Maj_Hi, and Maj_N, formulated previously in chapter 3, are listed in Table 5-9 using the settings below:

- Algorithm: decision tree: C4.5 (J48 in Weka) [48]
- Features: all features are used (features: 46)
- Dataset: the whole dataset is used (instances: 199)
- Validation procedure: 10 cross validation.
- Features selection algorithm: no features selection algorithms are used.

Table 5-9 Credibility classification results using simple majority voting method

Simple majority voting model	Dataset	Labelling include all crowd labellers, no experts Accuracy Rate
Maj_Class2	1:108, 2:50, 3:41	51.2563 %
Maj_Low	1:126, 2:32, 3:41	59.799 %
Maj_Hi	1:108, 2:35, 3:55	53.7688 %
Maj_N	1:109. 2:49, 3:41	55.2764 %

Detailed statistics for the accuracy results obtained using Maj_Class2 and Maj_Hi options is listed below in Table 5-10 and the complete classifier outputs is listed in Appendix C, Table C-6. Only these options are listed as Maj_Class2 is the more logical option to assign the tweet message credibility with "questionable" class when no majority voting is reached whereas for Maj_Hi, it was the most agreed vector with experts.

Table 5-10 Detailed statistics for the accuracy results obtained using Maj_Class2 and Maj_Hi

Maj_Class2

```
Correctly Classified Instances           102               51.2563 %
Incorrectly Classified Instances         97                48.7437 %
Kappa statistic                          0.1713
Mean absolute error                      0.3385
Root mean squared error                  0.5304
Relative absolute error                  84.4833 %
Root relative squared error              118.6085 %
Coverage of cases (0.95 level)           67.3367 %
Mean rel. region size (0.95 level)       55.4439 %
Total Number of Instances                199

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall    F-Measure   MCC       ROC Area   PRC
Area  Class
                 0.713     0.429     0.664       0.713     0.688       0.287     0.658      0.625
1
                 0.360     0.161     0.429       0.360     0.391       0.211     0.614      0.374
2
                 0.171     0.215     0.171       0.171     0.171       -0.044    0.419      0.190
3

Weighted Avg.    0.513     0.317     0.503       0.513     0.507       0.200     0.598      0.472

=== Confusion Matrix ===

  a  b  c   <-- classified as
 77 11 20 |  a = 1
 18 18 14 |  b = 2
 21 13  7 |  c = 3
```

Maj_Hi

```
Correctly Classified Instances           107               53.7688 %
Incorrectly Classified Instances         92                46.2312 %
Kappa statistic                          0.2192
Mean absolute error                      0.3321
Root mean squared error                  0.5199
Relative absolute error                  83.3385 %
Root relative squared error              116.5693 %
Coverage of cases (0.95 level)           68.8442 %
Mean rel. region size (0.95 level)       60.804  %
Total Number of Instances                199

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall    F-Measure   MCC       ROC Area   PRC
Area  Class
                 0.694     0.374     0.688       0.694     0.691       0.321     0.649      0.612
1
                 0.278     0.123     0.333       0.278     0.303       0.167     0.597      0.258
2
                 0.400     0.264     0.367       0.400     0.383       0.133     0.531      0.326
3
Weighted Avg.    0.538     0.298     0.535       0.538     0.536       0.241     0.607      0.469
```

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 75  9 24 |  a = 1
 12 10 14 |  b = 2
 22 11 22 |  c = 3
```

5.3.2.2 Classification results after applying selective proposed weighting measures

Using labelling obtained by similarity and accuracy proposed measures, calculated previously in chapter 4, Table 5-11 shows the accuracy results of credibility classification using the settings below:

- Algorithm: decision tree: C4.5 (J48 in Weka) [48]
- Features: all features are used (features: 46)
- Dataset: the whole dataset is used (instances: 199)
- Validation procedure: 10 cross validation.
- Features selection algorithm: no features selection algorithms are used.

Table 5-11 Credibility classification results using selective proposed measures

Selected labellers' weighting measures	**Dataset**	**Labelling include all crowd labellers, no experts Accuracy Rate**
Similarity		
Cosine Similarity Algorithm	1:111, 2:35, 3:53	55.78 %
Jaccard Similarity Algorithm	1:120, 2:35, 3:44	56.78 %
Accuracy		
Variance Accuracy	1:114, 2:39, 3:46	60.80 %
Standard Deviation Accuracy	1:110, 2:43, 3:46	55.28 %

Detailed statistics for the accuracy results obtained using labellers' weighting with similarity measure is listed below in Table 5-12, followed by the statistics results from the accuracy measure in Table 5-13. A complete classifier outputs is listed in Appendix C, Table C-7.

Table 5-12 Detailed statistics for the accuracy results obtained using similarity measures

Cosine similarity algorithm

```
Correctly Classified Instances          111               55.7789 %
Incorrectly Classified Instances        88                44.2211 %
Kappa statistic                         0.2258
Mean absolute error                     0.3137
Root mean squared error                 0.505
Relative absolute error                 79.9665 %
Root relative squared error             114.1285 %
Coverage of cases (0.95 level)          71.3568 %
Mean rel. region size (0.95 level)      57.7889 %
Total Number of Instances               199

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area  PRC Area
Class
                 0.721     0.443     0.672       0.721    0.696       0.281    0.632     0.627
1
                 0.314     0.091     0.423       0.314    0.361       0.252    0.660     0.299
2
                 0.377     0.233     0.370       0.377    0.374       0.144    0.538     0.330
3
Weighted Avg.    0.558     0.325     0.548       0.558    0.551       0.239    0.612     0.490
```

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 80  8 23 |  a = 1
 13 11 11 |  b = 2
 26  7 20 |  c = 3
```

Jaccard similarity algorithm

```
Correctly Classified Instances           113               56.7839 %
Incorrectly Classified Instances         86                43.2161 %
Kappa statistic                          0.1862
Mean absolute error                      0.3067
Root mean squared error                  0.4933
Relative absolute error                  82.3644 %
Root relative squared error              114.5078 %
Coverage of cases (0.95 level)           73.8693 %
Mean rel. region size (0.95 level)       59.9665 %
Total Number of Instances                199

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 0.767    0.506    0.697      0.767    0.730      0.270    0.665     0.696
1
                 0.257    0.091    0.375      0.257    0.305      0.194    0.549     0.258
2
                 0.273    0.200    0.279      0.273    0.276      0.073    0.549     0.272
3
Weighted Avg.    0.568    0.366    0.548      0.568    0.555      0.213    0.619     0.525

=== Confusion Matrix ===

  a  b  c   <-- classified as
 92  9 19 |  a = 1
 14  9 12 |  b = 2
 26  6 12 |  c = 3
```

Table 5-13 Detailed statistics for the accuracy results obtained using accuracy measures

Variance accuracy

```
Correctly Classified Instances           121               60.804  %
Incorrectly Classified Instances         78                39.196  %
Kappa statistic                          0.2932
Mean absolute error                      0.2788
Root mean squared error                  0.4799
Relative absolute error                  71.902  %
Root relative squared error              109.118  %
Coverage of cases (0.95 level)           73.8693 %
Mean rel. region size (0.95 level)       55.4439 %
Total Number of Instances                199

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 0.816    0.412    0.727      0.816    0.769      0.417    0.716     0.695
1
                 0.333    0.113    0.419      0.333    0.371      0.242    0.606     0.315
2
                 0.326    0.163    0.375      0.326    0.349      0.171    0.578     0.310
3
Weighted Avg.    0.608    0.296    0.585      0.608    0.594      0.326    0.662     0.531

=== Confusion Matrix ===

  a  b  c   <-- classified as
 93  7 14 |  a = 1
 15 13 11 |  b = 2
 20 11 15 |  c = 3
```

Standard deviation accuracy

```
Correctly Classified Instances           110               55.2764 %
Incorrectly Classified Instances         89                44.7236 %
Kappa statistic                          0.2541
Mean absolute error                      0.3222
Root mean squared error                  0.5173
Relative absolute error                  81.1353 %
Root relative squared error              116.1846 %
Coverage of cases (0.95 level)           66.8342 %
Mean rel. region size (0.95 level)       55.7789 %
Total Number of Instances                199
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
                 0.682    0.360    0.701      0.682    0.691      0.321    0.628     0.621
1
                 0.395    0.199    0.354      0.395    0.374      0.189    0.538     0.273
2
                 0.391    0.170    0.409      0.391    0.400      0.225    0.600     0.307
3
Weighted Avg.    0.553    0.281    0.559      0.553    0.555      0.271    0.602     0.473


=== Confusion Matrix ===

  a  b  c   <-- classified as
 75 18 17 |  a = 1
 17 17  9 |  b = 2
 15 13 18 |  c = 3
```

5.3.2.3 Classification results after applying the proposed weighting aggregation model

Using labelled dataset constructed by combining labellers' weights from different measures (similarity, accuracy, agreement, and majority consensus), Table 5-14 shows credibility classification accuracy results for this proposed model using the settings below:

- Algorithm: decision tree: C4.5 (J48 in Weka) [48]
- Features: different options are applied to investigate the usefulness of proposed features.
    - All features are used (features: 46).
    - Content features only.
    - Author features only.
    - Using selective features based on features selection algorithms: The extraction of features yields more than 45 features. In order to discard irrelevant features (which may degrade classification performance), we carried out a feature selection before classification. We decided to use different famous feature selection algorithms from Weka. There are many algorithms suggested in the literature for selection of a subset of features. We tested the following algorithms: Chi square (X2) [98], Correlation Attribute [99], and Relief algorithm [100], on the Ranker algorithm that ranks features according to average merit and average rank.
- Dataset: the whole dataset is used (Instances: 199) (1:113, 2:35, 3:51)
- Validation procedure: 10 cross validation.

Table 5-14 Credibility classification results using proposed weighting aggregation model

Features used	Feature selection algorithm	Accuracy Rate
All content and author features – 46 features	NA	58.794 %
Only content features – 24 features	NA	61.3065 %
Only author features – 22 features	NA	55.2764 %
13 selective ranked features: topic, author_ratiofwfl, author_flrno, author_logflrno, author_verif, author_hasweb, content_rank, content_hasURL, content_haslaugh, content_squest, content_pronnoT, content_pronounsD, content_hasimage	Chi Square (X2)	63.3166 %
9 selective ranked features: topic, author_logflrno, author_verif, author_hasweb, author_flrno, content_rank, content_pronnoT, content_hasimage, author_yearsntwt	Correlation Attribute	64.3216 %
14 selective ranked features: topic, content_hasimage, author_descrelate, content_hasURL, author_locrelate, author_emp, author_hasweb, content_wordsno, content_alldelicate, content_hashno, author_allinf1, author_allrelate, author_logflrno, author_verif	Relief Algorithm	65.8291 %

Table 5-15 shows the experimental results using the proposed weighting aggregation approach, clearly it shows that the classification rate of the proposed model (59% - 66%) is more desirable than the classification rate of simple majority voting approach using both the logical option Maj_Class2: 51.26% and the most agreed vector with experts: Maj_Hi: 53.77%. A detailed statistics data for the accuracy of our proposed method is listed below in Table 5-15 using original feature set and Relief Algorithm feature set. A complete classifier outputs is listed in Appendix C, Table C-8.

Table 5-15 Detailed statistics for the accuracy results obtained using proposed weighting aggregation model

Weighting aggregation model - All features

```
Correctly Classified Instances          117               58.794  %
Incorrectly Classified Instances        82                41.206  %
Kappa statistic                         0.281
Mean absolute error                     0.301
Root mean squared error                 0.4968
Relative absolute error                 77.52   %
Root relative squared error             112.8718 %
Coverage of cases (0.95 level)          71.3568 %
Mean rel. region size (0.95 level)      59.6315 %
Total Number of Instances               199

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall    F-Measure   MCC       ROC Area   PRC
Area  Class
                 0.743     0.384     0.718       0.743     0.730       0.362     0.689      0.670
1
                 0.314     0.122     0.355       0.314     0.333       0.202     0.528      0.232
2
                 0.431     0.196     0.431       0.431     0.431       0.235     0.556      0.349
3
Weighted Avg.    0.588     0.290     0.581       0.588     0.584       0.301     0.627      0.511

=== Confusion Matrix ===

  a  b  c   <-- classified as
 84 11 18 |  a = 1
 13 11 11 |  b = 2
 20  9 22 |  c = 3
```

Weighting aggregation model - Relief Algorithm feature set

```
Correctly Classified Instances          131               65.8291 %
Incorrectly Classified Instances        68                34.1709 %
Kappa statistic                         0.3891
Mean absolute error                     0.2749
Root mean squared error                 0.4355
Relative absolute error                 70.7865 %
Root relative squared error             98.9265 %
Coverage of cases (0.95 level)          82.9146 %
Mean rel. region size (0.95 level)      74.3719 %
Total Number of Instances               199

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall    F-Measure   MCC        ROC Area   PRC
Area  Class
                 0.823     0.372     0.744       0.823     0.782       0.462      0.751      0.737
1
                 0.429     0.079     0.536       0.429     0.476       0.382      0.678      0.366
2
                 0.451     0.155     0.500       0.451     0.474       0.306      0.623      0.416
3
Weighted Avg.    0.658     0.265     0.645       0.658     0.649       0.408      0.705      0.590

=== Confusion Matrix ===

  a  b  c   <-- classified as
 93  6 14 |  a = 1
 11 15  9 |  b = 2
 21  7 23 |  c = 3
```

To validate the effectiveness of our proposed credibility model, we compare it with existing Arabic credibility assessment studied by Al-Eidan, Al-Khalifa, & Al-Salman 2010 [9] and Al-Khalifa & Al-Eidan 2011 [8]. A comparison based on number of used features and classification results is listed in Table 5-16.

Table 5-16 A comparison of the proposed model with existing Arabic model

Research work/ Criteria	**Al-Eidan, Al-Khalifa, & Al-Salman 2010 [9] and Al-Khalifa & Al-Eidan 2011 [8]**	**Weighting aggregation model**
Features	**3 Content features:** Similarity with verified content , inappropriate Words, Linking to authoritative News Sources **2 Author features:** Verified, Author overall degree from (TwitterGrader.com) **Total : 5 features**	**24 Content features** **22 Author features** **Total: 46 features** (refer to Table 5-2 for complete features listing)
Classification Results	**Using similarity alone features:** The average precision value : 0.52 The average recall value is 0.56 **Using all features:** The average precision value : 0.48 The average recall value is 0.56	**Using Relief Algorithm feature set (14 selective ranked features):** The average precision value : 0.645 The average recall value is 0.658 F-Measure: 0.649 **Using all features:** The average precision value : 0.581 The average recall value is 0.588 F-Measure: 0.584

5.3.2.3.1 Classification results using different classification algorithms

We also compared the performance of our proposed model using the following classification algorithms: C4.5 (J48 in Weka) [48], Random Forest tree [101], Naive Bayes [102], Logistic Regression, SVM (SMO in Weka) and k-Nearest Neighbour (IBk in Weka), as implemented in Weka 3.6.10. Table 5-17 below shows the classification accuracy results using only two credibility classes {1, 3} as some of the used classifiers (e.g. SVM) are binary classifiers.

Table 5-17 Credibility classification results using different classification algorithms

Classification model	Classification algorithm	Dataset	Labelling include all crowd labellers, no experts Accuracy Rate
Decision tree Algorithms	C4.5 (J48)	1:113, 3:51	67.0732 %
	Random Forest	1:113, 3:51	73.7805 %
Bayesian Algorithms	Naïve Bayes	1:113, 3:51	59.1463 %
Regression Algorithms	Simple Logistic Regression	1:113, 3:51	68.9024 %
SVM	SVM (SMO)	1:113, 3:51	74.3902 %
Instance-based Algorithms	k-Nearest Neighbour (IBk)	1:113, 3:51	67.6829 %

A detailed statistics data for the classification accuracy of our proposed method using only two credibility classes is listed below in Table 5-18 based on different classification algorithms whereas a complete classifier outputs are listed in Appendix C, Table C-9.

Table 5-18 Detailed statistics data for the accuracy results using different classification algorithms

Weighting aggregation model – C4.5 (J48) decision tree

```
Correctly Classified Instances           110               67.0732 %
Incorrectly Classified Instances         54                32.9268 %
Kappa statistic                          0.2398
Mean absolute error                      0.3457
Root mean squared error                  0.5518
Relative absolute error                  80.4864 %
Root relative squared error              119.1857 %
Coverage of cases (0.95 level)           76.8293 %
Mean rel. region size (0.95 level)       67.0732 %
Total Number of Instances                164

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall    F-Measure   MCC       ROC Area   PRC
Area  Class
                 0.752     0.510     0.766       0.752     0.759       0.240     0.586      0.705
1
                 0.490     0.248     0.472       0.490     0.481       0.240     0.586      0.425
3
Weighted Avg.    0.671     0.428     0.674       0.671     0.672       0.240     0.586      0.618

=== Confusion Matrix ===

  a  b   <-- classified as
 85 28 |  a = 1
 26 25 |  b = 3
```

Weighting aggregation model – Random Forest tree

```
Correctly Classified Instances           121               73.7805 %
Incorrectly Classified Instances         43                26.2195 %
Kappa statistic                          0.2931
Mean absolute error                      0.3581
Root mean squared error                  0.4199
Relative absolute error                  83.3667 %
Root relative squared error              90.7011 %
Coverage of cases (0.95 level)           100      %
Mean rel. region size (0.95 level)       95.7317 %
Total Number of Instances                164
```

```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall    F-Measure   MCC       ROC Area   PRC
Area  Class
                 0.920     0.667     0.754       0.920     0.829       0.322     0.749      0.865
1
                 0.333     0.080     0.654       0.333     0.442       0.322     0.749      0.609
3
Weighted Avg.    0.738     0.484     0.723       0.738     0.708       0.322     0.749      0.786

=== Confusion Matrix ===

   a   b   <-- classified as
 104   9 |   a = 1
  34  17 |   b = 3
```

Weighting aggregation model – Naïve Bayes

```
Correctly Classified Instances          97               59.1463 %
Incorrectly Classified Instances        67               40.8537 %
Kappa statistic                         0.1977
Mean absolute error                     0.4136
Root mean squared error                 0.5845
Relative absolute error                 96.2762 %
Root relative squared error             126.2429 %
Coverage of cases (0.95 level)          80.4878 %
Mean rel. region size (0.95 level)      71.0366 %
Total Number of Instances               164

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall    F-Measure   MCC       ROC Area   PRC
Area  Class
                 0.549     0.314     0.795       0.549     0.649       0.218     0.636      0.758
1
                 0.686     0.451     0.407       0.686     0.511       0.218     0.637      0.417
3
Weighted Avg.    0.591     0.357     0.674       0.591     0.606       0.218     0.636      0.652

=== Confusion Matrix ===

  a  b   <-- classified as
 62 51 |  a = 1
 16 35 |  b = 3
```

Weighting aggregation model – Simple Logistic Regression

```
Correctly Classified Instances          113              68.9024 %
Incorrectly Classified Instances        51               31.0976 %
Kappa statistic                         0.2106
Mean absolute error                     0.3899
Root mean squared error                 0.4684
Relative absolute error                 90.7768 %
Root relative squared error             101.1821 %
Coverage of cases (0.95 level)          96.9512 %
Mean rel. region size (0.95 level)      94.8171 %
Total Number of Instances               164

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall    F-Measure   MCC       ROC Area   PRC
Area  Class
                 0.841     0.647     0.742       0.841     0.788       0.217     0.636      0.760
1
                 0.353     0.159     0.500       0.353     0.414       0.217     0.636      0.444
3
Weighted Avg.    0.689     0.495     0.667       0.689     0.672       0.217     0.636      0.662


=== Confusion Matrix ===

  a  b   <-- classified as
 95 18 |  a = 1
 33 18 |  b = 3
```

Weighting aggregation model – SVM (SMO)

```
Correctly Classified Instances          122              74.3902 %
Incorrectly Classified Instances        42               25.6098 %
Kappa statistic                         0.3537
Mean absolute error                     0.2561
Root mean squared error                 0.5061
Relative absolute error                 59.6186 %
Root relative squared error             109.3073 %
Coverage of cases (0.95 level)          74.3902 %
Mean rel. region size (0.95 level)      50      %
Total Number of Instances               164

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall    F-Measure   MCC       ROC Area   PRC
Area  Class
                 0.876     0.549     0.780       0.876     0.825       0.362     0.664      0.768
1
                 0.451     0.124     0.622       0.451     0.523       0.362     0.664      0.451
3
Weighted Avg.    0.744     0.417     0.730       0.744     0.731       0.362     0.664      0.670
```

```
=== Confusion Matrix ===

  a  b   <-- classified as
 99 14 |  a = 1
 28 23 |  b = 3
```

Weighting aggregation model – k-Nearest Neighbour (IBk)

```
Correctly Classified Instances         111               67.6829 %
Incorrectly Classified Instances       53                32.3171 %
Kappa statistic                        0.2335
Mean absolute error                    0.3255
Root mean squared error                0.5647
Relative absolute error                75.7833 %
Root relative squared error            121.9747 %
Coverage of cases (0.95 level)         67.6829 %
Mean rel. region size (0.95 level)     50      %
Total Number of Instances              164

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall    F-Measure   MCC       ROC Area   PRC
Area  Class
                 0.779     0.549     0.759       0.779     0.769       0.234     0.621      0.750
1
                 0.451     0.221     0.479       0.451     0.465       0.234     0.621      0.405
3
Weighted Avg.    0.677     0.447     0.672       0.677     0.674       0.234     0.621      0.643

=== Confusion Matrix ===

  a  b   <-- classified as
 88 25 |  a = 1
 28 23 |  b = 3
```

The experimental classification results for two classes {1, 3} using labelling from proposed aggregation model yield (59.1463% - 74.3902%) based on different classification algorithms. By introducing the Relief feature selection algorithm to our dataset using the top ranked 12 features (content_hasimage, topic, content_hasURL, author_locrelate, author_descrelate, author_emp, content_alldelicate, content_wordsno, author_hasweb, author_logflrno), the classification result using Random forest tree reached 77.439 % which is considered a promising result. A detailed statistics data for the classification accuracy is listed below in Table 5-19.

Table 5-19 Detailed statistics for the accuracy results using Random forest tree algorithm

Weighting aggregation model – Random Forest tree - Relief Algorithm feature set

```
Correctly Classified Instances          127               77.439  %
Incorrectly Classified Instances        37                22.561  %
Kappa statistic                         0.4404
Mean absolute error                     0.3184
Root mean squared error                 0.4055
Relative absolute error                 74.1258 %
Root relative squared error             87.5846 %
Coverage of cases (0.95 level)          99.3902 %
Mean rel. region size (0.95 level)      91.7683 %
Total Number of Instances               164

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall    F-Measure   MCC       ROC Area   PRC
Area  Class
                 0.885     0.471     0.806       0.885     0.844       0.447     0.804      0.901
1
                 0.529     0.115     0.675       0.529     0.593       0.447     0.804      0.625
3
Weighted Avg.    0.774     0.360     0.766       0.774     0.766       0.447     0.804      0.815

=== Confusion Matrix ===

   a   b   <-- classified as
 100  13 |   a = 1
  24  27 |   b = 3
```

5.3.3 Effect of majority voting level on classification Accuracy

In order to understand the importance of having a labelled dataset with a high level of agreement, this section studied the effect of majority voting level on classification accuracy. Our goal here was to explore whether the level of majority voting correlates with learning performance. Table 5-20 shows the credibility prediction accuracy results under two level of majority voting ratio: low and high. Specifically, we divided the dataset into two subsets; the first subset consists of all tweets combined with their features where the labelling by majority voting is high; the ratio of labels with majority voting is greater than 50% compared with other class labels; more than half of the labellers agree on the same labelling. The other half of the dataset consists of tweet messages with their features where majority voting less than 50% which means less than half of the labellers agree on the same labels. Table 5-20 shows the classification accuracy for both subsets and it demonstrates that dataset with less noisy labels along with a higher level agreement between labellers can achieve better classification accuracy results. With labelled dataset having low percentage of majority voting class, the accuracy was in the range (32% – 50.5%) whereas labelled dataset with high percentage of majority voting class, was between (62.8% - 66.6%). Of course, the higher majority level, the higher the quality of the training data resulting in the better the performance of the learned model. This experiment defined the research problem clearly and demonstrated the significance of improving the quality of labelling by reducing the effect of noisy labels. In case of examining a labelled dataset, having 52% as an average percentage for the majority voting class, and where no majority agreement obtained for almost 11.56% of the dataset (as in our dataset case), we argue that introducing this proposed labellers' weighting model has improved the quality of dataset and resolve the problem of conflicts judgments.

Table 5-20 Majority voting level versus classification accuracy

Labelling include all crowd labellers, no experts			Labelling include all labellers, with experts		
Percentage of majority voting class	**Dataset**	**Accuracy Rate**	**Percentage of majority voting class**	**Dataset**	**Accuracy Rate**
High percentage of majority voting class (>50%) 94 Instances	1:67, 2:9, 3:18	62.766 %	High percentage of majority voting class (>50%) 96 Instances	1:70, 2:9, 3:17	66.6667 %
Low percentage of majority voting class (<=50%) 105 Instances	Maj_Class2 1:41, 2:41, 3:23	44.7619 %	Low percentage of majority voting class (<=50%) 103 Instances	Maj_Class2 1:37, 2:44, 3:22	45.6311 %
	Maj_Low: 1:59:2:23,3:23	50.4762 %		Maj_Low: 1:58:2:23,3:22	46.6019 %
	Maj_Hi 1:41, 2:27, 3:37	45.7143 %		Maj_Hi: 1:37, 2:27, 3:39	32.0388 %
	Maj_N 1:42, 2:40, 3:23	40.9524 %		Maj_N: 1:38, 2:43, 3:22	41.7476 %

A complete classifier output listings of the classification results using all labellers for all majority voting ratio options is found in Appendix C, Table C-10.

5.4 Conclusions from Credibility Detection Using Feature-based Approaches

This chapter is devoted to the last two stages for credibility detection illustrated in the proposed system architecture which are: features extraction and credibility classification. In features extraction and analysis step, messages are represented by a set of measured features. A total of 46 features were extracted and computed in this study, which contain 24 content features and 22 author features. Later on this chapter, we used these features to detect and classify credibility using both: 1) a statistical approach based on features frequencies and 2) machine learning algorithms.

Overall, there are common-shared prominent features from both the survey results and the statistical analysis model. It was found that features related to the source authority and expertise, and data quality factors are common in both methods and would be used to distinguish low-credibility messages from other high-credibility messages. Furthermore, messages with low data quality content lead to a low credibility perception, features such as the availability of swearing/ inappropriate, laughing words, informal dialectal words, spelling errors, exclamation marks, emoticons, and pronouns might be indicators for low-credibility messages. Yet, features which are related to the authority of the source factor were extensively employed

to identify higher credibility messages. Indeed, features such as having “verified” account, large number of followers, favourites and tweets, a webpage, an old Twitter account, and a bio related to the discussed topic, all appeared to be signs for higher credibility. Surprising findings involved authors’ bio information; the statistical model suggests that having a higher degree of authors’ self-disclosure such as stating one’s education level and employment did not affect positively in the believability of messages. Our findings attested to the importance of author features which also consider a metadata-based features meaning that is more robust and consistent for determining credibility as it is language independence. As for content features, it is heavily dependent on content language and extracting them is comparatively lower than author features as they need special text analysers’ tools.

By studying the relation between the similarity and agreement between labellers and the credibility used features, it was evident that in general, having similar labellers does not mean assigning credibility judgments using the same features. In addition, by using alpha agreement measure, the most agreed labellers show almost 100% features similarity for the low-credibility class. A sound explanation for this finding is that the majority of labels were within low-credibility class which make it more reasonable to reach features similarity within this class.

With regards to proposed model classification performance, the experimental results listed in the last sections consider acceptable compared with results obtained by simple majority voting method. The classification accuracy rates of the simple majority voting approach was (51.26% and 53.77%) for the logical option (Maj_Class2) and the most agreed vector with experts (Maj_Hi), respectively. On the other hand, classification accuracy rate of the proposed model was 59% using decision tree classifier and utilizing all features, and reached to 66% for selected features from both content and author. In addition, the experimental results show that proposed model has a good performance (average precision: 0.645, average recall: 0.658) using Relief selection algorithm feature set (14 selective ranked features), comparable to existing Arabic credibility model using similarity alone features (average precision: 0.52, average recall: 0.56). Better results were achieved using two credibility classes {1, 3} where the accuracy results reached 77.4% using Random forest tree classifier with a Relief selection algorithm feature set. Hence, we assume one of reasons for comparatively low classification results regarding the high-credibility class {3}, is the small sample size from the class which did not allow for in-depth learning.

The last section of this chapter covers an experiment to study the relation between the quality of the dataset and the performance of the classifier. It demonstrates that dataset with less noisy labels - higher agreement level between labellers can achieve better classification accuracy results. Moreover, using labelled dataset with low level of agreement between labellers means low ratio of majority voting class, the accuracy was in the range (32% – 50%) whereas the labelled dataset with high percentage of majority voting class, it was between (63.8% - 66.7%). This finding clarifies the research significance and that improving the quality of labelling by reducing the effect of noisy labels would yield better classification results.

Based on the results presented, we recommend the following processing pipeline to build credibility assessment model using datasets with noisy labels (disagreed judging credibility scores). First, a set of labellers' weights vectors is created based on measures that capture similarity, accuracy, agreement, majority consensus, and propensity to trust factors. A detailed description of the computed measures is given in chapter 4 – section 4.1. Then, we aggregate such estimated multiple weights in order to create the final ranks of evaluated labellers. Accordingly, final dataset credibility labelling is constructed based on labellers' aggregated weights. It was evident that the inferred labelling using "labellers' weighting aggregation model" aligns more with the experts' ratings compared to the common majority voting method. Finally, a decision tree J48 classifier is trained on constructed labels to predict credibility of tweet messages. Classification rate accuracy 65.8291 % with an average precision: 0.645, and average recall: 0.658 has been achieved using Relief selection algorithm feature set (14 selective ranked features). Better results were achieved using two credibility classes {1, 3} where the accuracy results reached 77.4% using Random forest tree classifier with a Relief selection algorithm feature set.

Noted that the recommended pipeline: labellers' weighting aggregation model, which based on evaluating the reliability of crowd labellers in order to determine objective labels, may also work for research required opinion labelling where disagreements (noising) are expected.

6 Conclusions and Future work

Along with the continued dependence on different UGC platforms specifically the microblogging medium as an information source for Arab users, comes a continued need for research addressing Arabic content credibility. Yet, research in the area of Arabic content credibility is minimal in comparison to studies in English and other languages. This study presented in-depth research of credibility assessment methods in order to detect credibility for Arabic tweet messages in the presence of disagreed judging credibility scores. The main problem discussed in this study was how to construct correct labels "ground truth" that can be used to build credibility model in light of noisy subjective judgments. Therefore, this study proposes to evaluate the reliability of crowd labellers in order to determine objective credibility labels. This study evaluated indicators of crowd labellers encapsulating two general concepts: 1) evaluating labellers based on their characteristics collected in the user study at the time of credibility evaluation such as labellers' Twitter usage, and topic familiarity. 2) evaluating labellers based on the quality of their credibility ratings, insofar as how labellers' weighting are derived from the ratings evaluation itself which is the primary focus of this research.

To achieve the stated goal, the study was divided into three main phases: in the first stage, we created Twitter dataset posts from different topics, accompanied by human annotation. We employed the idea of crowdsourcing where users could explicitly express their opinions about credibility of a set of tweets coupled with their features. One of the outcomes of this study stage is a corpus of human annotated Arabic messages along with computed features that could be used for further research. We distinguished three main groups of features: authority and topical expertise (of the source), data quality (of the content), and popularity (of the content and the source). In addition to this, we investigated the level of agreement between experts' ratings and the crowd in order to identify the expert who best represents the crowd. This technique allowed us to select the most representative expert on a needs basis.

- It is concluded from the analysis of submitted credibility rating values and the collected labellers' data covered by this stage of the study that Arab users are more sceptical about believing Twitter online information. Most of the credibility labels gathered were among the low-credibility class. In addition, as expected, and in accordance with other studies' results, labellers' data and method of presentation have a slight impact on the perception of credibility. The following elements slightly influence the perception of credibility:

- Crowd labellers who used Twitter-presentation for credibility annotations mostly submitted low scores whereas other labellers who used text-presentation, scores varied between all classes.
- In regards to labellers' characteristics and its effect on credibility perception, the study suggested that Arab youth under 35 years old have relatively higher trust in Twitter information than older adults which might match results with other studies that revealed the younger generation [4] is more influenced by online information.
- With respect to labellers' gender factor, males were harsher in their credibility judgments as they gave less high-credibility scores to tweet messages. This finding also matched results by Fogg et al. 2001 [26] where men, in general, assigned lower credibility ratings.
- Generally, labellers from different education levels perceived credibility almost the same way with slight inclination to assign more low credibility scores as labellers have higher levels of education.
- Data collected regarding labellers' Twitter features and frequency of Twitter use confirmed that labellers who tend to use Twitter more often and have more influential Twitter features are inclined to trust Twitter content as they assigned more high-credibility ratings.
- As part of analysing personality labellers' features, individuals with high trusting characteristics offered higher credibility evaluations compared to other labellers.
- Labellers with higher topic familiarity and interest assigned more low-credible rating scores.

- Apart from the annotations agreement calculations and interpretation sections which also covered in this phase, the study revealed the following results:
  - There is a "slight" agreement between labellers on assessing the credibility of tweets. This can be interpreted as due credibility judgments labelling is after all "opinion labelling" which prone to high subjectivity.
  - The agreement level of the non-expert crowd outperformed experts' annotations agreement value. This evidence suggested that disagreements among individual labellers (including experts) might arise because of their inherent biases, expertise and propensity to trust levels, but introducing a larger population sample in the annotation task, might reduce the tendency towards bias. Wisdom-of-the-crowds

research addressed this issue and indicated the advantage of crowd over individual judgments in reducing the effect of individualism bias [103].

- The inter-labellers agreement values were affected by the number of labellers, and to have stable agreement values, more labellers' annotations (at least the contributions from 15 labellers) should be included. It should be noted that combining more multiple diversity labels within may have the effect of reducing labeller bias and thus improving the quality of the data.

The second stage of the study concentrates on the second concept which is dedicated to the proposed credibility model. This section investigates the possibility of determining the credibility of Arabic tweet messages given conflicted credibility labels. The proposed solution mechanism is based on deducing the correct credibility labels of the tweet messages by analysing and estimating labellers' reliability to justify the quality of their credibility ratings. The proposed model uses different criteria measurements for evaluating the reliability of labellers to deliberately reduce the influence of unreliable crowd labellers. To evaluate labellers' reliability weights, the basic traits of each labeller are extracted explicitly from the user survey along with his/her labelling scores. Then, labeller's credibility scores for the tweet messages are used as inputs to generate labellers' reliability weights using mainly accuracy and similarity measurements. The proposed model then, manages the computed labellers' reliability weights to construct the correct credibility labels for the tweet messages.

In this proposed framework, we applied different measurements to weight the labellers and conducted experiments to assess how the proposed techniques might enhance the fairness and accuracy of the applied dataset and reduce the spontaneity of judgments. In order to evaluate the proposed model, we compare the labelling obtained from the expert labellers and those from the non-expert crowd labellers after applying the weighting mechanism. Using Krippendorff's alpha agreement measure, we find that similarity and accuracy weighting measures presented reasonable promising results and outperform the agreement values using simple majority voting. A "substantial" agreement of value 0.71 with expert labelling was attained for some of the proposed weighting methods. This result suggested that weighting labellers based on their contributions using multiple measures would improve the resulting crowd labelling quality similar to experts' labelling. Therefore, we recommend with this type of

labelling tasks (opinion labelling) which tends to be noisy by nature, a framework of multiple labellers' reliability measures which might yield overall higher quality credibility judgments.

The final stage of the thesis consecrates on the use of the dataset given computed features and constructed labels with feature-based approaches mainly: relative features frequencies and decision tree algorithm to detect the credibility of messages. In this stage, we thoroughly evaluated various state-of-the-art features and reported results in two features-based approaches to detect credibility. Using the statistical approach, both implicit and explicit methods have been used to check the prominent features consumed to assess the tweet messages' credibility. In support of the implicit method, an experimental study was conducted to statistically examine and compare how the features were distributed within the annotated dataset. A histogram was used to detect the percentage of occurrences of computed features in different credibility classes. On behalf of the explicit method, a user survey was used to rate the importance of features on assessing messages' credibility. On the whole, there were common-shared prominent features from both: the survey results and the statistical analysis model. It was found that features related to the source authority and expertise, and data quality factors were common in both methods hence could be used to identify high-credibility messages. Our findings also propose the importance of author features which also consider metadata-based features, meaning that is more robust and consistent when determining credibility as it is language independent. In terms of content features which are heavily dependent on content data and used language, it was noteworthy that extracting is comparatively lower than author features as content features need intelligent lexical analysers' tools. This underscores the importance to verify how far the research has progressed for text analysis, particularly with respect to the Arabic language.

Regarding machine learning classifier performance, the classification results of the proposed model presented in the last part of the thesis are considered to be acceptable compared with the ones obtained by simple majority voting method. Specifically, classification accuracy reached 58.79% for three credibility classes using decision tree classifier and improved to 65.8% using selected features from both content and author whereas the classification rate of simple majority voting approach was 51.26%. In case of using two credibility classes {1, 3}, the accuracy results based on our proposed model reached 77.4% using Random forest tree

classifier with a Relief selection algorithm feature set. Hence, the combined experiment results revealed the following critical central outcomes:

- We argue that source authority and expertise in addition to content data quality factor based on content linguistic features is more important than content popularity in identifying credible messages. This finding does not agree with descriptions of the majority of previous research where the popularity of tweet messages plays a role feature.
- We also discovered a relationship between most agreed, similar users and the shared credibility features. In addition, it was interesting to observe that similar labellers do not always correlate with assigning credibility judgments harnessing the same features. Shared features between labellers primarily appear with the most agreed labellers using alpha agreement resulting in close to 100% features similarity for low-credibility.
- We concluded that the credibility model that applied features selection algorithms yields enhanced performance in deference to using the whole set of features. This was due to the fact that some features were irrelevant and may have hindered the performance.
- We explored the relationship between the level of majority agreement and machine learning performance, and discovered that if labellers conclude approximate same judgemental labelling, the machine learning algorithm would outperform and achieve better classification accuracy results.

Simply stated, it was evident that adding an extra step to evaluate labellers leads to a more robust labelled dataset similar to experts' labelling for building the credibility classification model. Despite the moderate classification results, we achieved the research goal of improving the quality of dataset in the task of Arabic content credibility assessment on Twitter. To the best of our knowledge, this is the first documented initiative to model credibility in Arabic settings using noisy labels. Moreover, this study advises that employing the proposed labellers' weighting model, would improve the quality of a given "noisy" labelled dataset, where no majority voting can be obtained, and resolves the problem of conflict judgments. The proposed baseline credibility model may also serve as a starting point for future research based on crowd opinions where disagreements (noising) are expected.

6.1 Future Work

During the research and writing stages of this thesis there were several areas that could not be fully investigated or implemented which would present interesting challenges and problems to explore further. Below is a list of focus areas that require deeper and broader analysis:

- We expect our experiments to continue including more features and larger dataset labelled with additional topics' experts. Another parallel corpus should be created and labelled in order to have a larger dataset that can be used for classification with machine learning algorithms. Future research will look also into examining features from Arabic tweet messages related to rumour topics verified by an official sources (no labelling needed) and compare it to the features from truthful confirmed events.
- Users may possess multiple readings of credibility depending on to the type of information that will be evaluated. For instance, consider a situation when a user evaluates health information. In this case, it appears that authority and topical expertise (of the source), might be more important than other factors. This is a motivation for future work which aims to find the relationship between dataset topics and the presence/absence of different features and whether different topics credibility perception rely on different features and factors.
- Investigating Individuals' perceived credibility for online messages for similar events between different locations and cultures - as an example comparing credibility perceptions among English speaking countries and Arab countries audiences [US/UK vs. Arabs] would present interesting challenges. This project could be combined with an experimental quantitative study that compares how selected features are being distributed surrounding different contextual dimensions: cultural, situational, and topical variations. The aim is to develop an understanding of how different communities (both in the U.S.A/U.K and in the Arab countries) consume credibility of online messages during different events in order to determine when specific features are useful in determining credibility.
- Due to the fact that we can't control the time of credibility labelling with time of posted tweet messages, the online surveys usually do not capture the immediate real-time credibility judgments for labellers based on information and circumstances that was available at that time. Therefore, it is suggested to explore the possibility to implement real-time labelling solutions that consider the real-time nature of Twitter.

- The influence of the number of labellers on the inter-labellers agreement values needs more investigation in order to have a clear indication with the minimum labellers are required to reach a stable agreement value. In addition, it is suggested to study the influence of labellers set permutation on the stability of alpha agreement values computed on any combination of 15 labellers (among 52) on the same annotated dataset. "*a coefficient for assessing the reliability of data must treat coders as interchangeable"* Krippendorff, 2004 [67].
- Another research direction that further extends the credibility model proposed in this thesis is to incorporate filtering technique. For instance, in attempt to improve the data quality, we identify labellers with low computed weights below a certain threshold and drop their contributions; then study the agreement level of labelling with experts and the classification accuracy using only the selected labellers.
- Develop a crowd labelling workflow for opinion labelling tasks which monitor labellers and advise them to work independently but according to some clearly fixed specified criteria and instructions. This mechanism might reduce the crowd disagreements (noising) in labelling opinion tasks and produces standard representative annotation.
- Integrate the probabilistic model using the expectation maximization algorithm (EM) for estimating credibility ratings and also to validate the obtained labels using our proposed method [104].
- As data quality factor was important to detect credibility, we assume that generating a lexicon with a set of dispute Arabic patterns that could indicate a tweet message is disputed as a valuable content feature [105]. This is based on the assumption of an event would be within low-credibility or questionable class if there are percentage of tweets containing phrases like: "no proof that", "mistakenly believe that", "no evidence that", "it is not true that", "it is not clear that", "it is unlikely that", "it is wrong that", "it is denied that", "misconception that", "myth that", "it is speculated that", "it is delusion that", مصدرغير موثوق - التحقق – خبر نفي – الشكوك –لا يوجد اثبات –لا يوجد دليل –شخص غير ثقة – اشاعة
- Incorporate a larger dataset which broadens the opportunity to explore the relation between the level of majority voting and the accuracy of classification model using different majority voting ratio levels. With respect to the nature of our dataset, this study only covered two levels of majority voting ratio: low (<=50) and high (>50) and it is recommended for future work, to study the classification accuracy results for multiple levels of majority voting.

7 References

[1] S. T. Moturu and H. Liu, “Quantifying the trustworthiness of social media content,” *Distrib. Parallel Databases*, vol. 29, no. 3, pp. 239–260, 2011.

[2] A. J. Flanagin and M. J. Metzger, “Perceptions of Internet information credibility,” *Journal. Mass Commun. Q.*, vol. 77, no. 3, pp. 515–540, 2000.

[3] B. Sharifi, M.-A. Hutton, and J. Kalita, “Summarizing microblogs automatically,” in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 685–688.

[4] V. J. Rideout, U. G. Foehr, and D. F. Roberts, “Generation M [superscript 2]: Media in the Lives of 8-to 18-Year-Olds.,” *Henry J Kais. Fam. Found.*, 2010.

[5] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a Social Network or a News Media?,” in *Proceedings of the 19th International Conference on World Wide Web*, New York, NY, USA, 2010, pp. 591–600.

[6] Á. Cuesta, D. F. Barrero, and M. D. R-Moreno, “A Descriptive Analysis of Twitter Activity in Spanish around Boston Terror Attacks,” in *Computational Collective Intelligence. Technologies and Applications*, Springer, 2013, pp. 631–640.

[7] A. A. AlMansour, L. Brankovic, and C. S. Iliopoulos, “Evaluation of Credibility Assessment for Microblogging: Models and Future Directions,” in *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business*, New York, NY, USA, 2014, p. 32:1–32:4.

[8] H. S. Al-Khalifa and R. M. Al-Eidan, “An experimental system for measuring the credibility of news content in Twitter,” *Int. J. Web Inf. Syst.*, vol. 7, no. 2, pp. 130–151, 2011.

[9] R. M. B. Al-Eidan, H. S. Al-Khalifa, and A. S. Al-Salman, “Measuring the credibility of Arabic text content in Twitter,” in *Digital Information Management (ICDIM), 2010 Fifth International Conference on*, 2010, pp. 285–291.

[10] B. Hilligoss and S. Y. Rieh, “Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context,” *Inf. Process. Manag.*, vol. 44, no. 4, pp. 1467–1484, 2008.

[11] B. J. Fogg, “Prominence-interpretation theory: Explaining how people assess credibility online,” in *CHI’03 extended abstracts on human factors in computing systems*, 2003, pp. 722–723.

[12] S. Bhattacharya, H. Tran, P. Srinivasan, and J. Suls, “Belief Surveillance with Twitter,” in *Proceedings of the 4th Annual ACM Web Science Conference*, New York, NY, USA, 2012, pp. 43–46.

[13] C. Castillo, M. Mendoza, and B. Poblete, “Information Credibility on Twitter,” in *Proceedings of the 20th International Conference on World Wide Web*, New York, NY, USA, 2011, pp. 675–684.

[14] A. Gupta and P. Kumaraguru, “Credibility Ranking of Tweets During High Impact Events,” in *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, New York, NY, USA, 2012, p. 2:2–2:8.

[15] B. Kang, J. O'Donovan, and T. Höllerer, "Modeling Topic Specific Credibility on Twitter," in *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, New York, NY, USA, 2012, pp. 179–188.

[16] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor Has It: Identifying Misinformation in Microblogs," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2011, pp. 1589–1599.

[17] X. Xia, X. Yang, C. Wu, S. Li, and L. Bao, "Information Credibility on Twitter in Emergency Situation," in *Proceedings of the 2012 Pacific Asia Conference on Intelligence and Security Informatics*, Berlin, Heidelberg, 2012, pp. 45–59.

[18] J. O'Donovan, B. Kang, G. Meyer, T. Hollerer, and S. Adalii, "Credibility in context: An analysis of feature distributions in twitter," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, 2012, pp. 293–301.

[19] L. Madlberger and A. Almansour, "Predictions based on Twitter—A critical view on the research process," in *Data and Software Engineering (ICODSE), 2014 International Conference on*, 2014, pp. 1–6.

[20] M. L. Jensen, P. B. Lowry, and J. L. Jenkins, "Effects of automated and participative decision support in computer-aided credibility assessment," *J. Manag. Inf. Syst.*, vol. 28, no. 1, pp. 201–234, 2011.

[21] M. Lynch, *The Arab uprising: The unfinished revolutions of the new Middle East*. PublicAffairs, 2013.

[22] C. I. Hovland and W. Weiss, "The influence of source credibility on communication effectiveness," *Public Opin. Q.*, vol. 15, no. 4, pp. 635–650, 1951.

[23] D. W. Stacks and M. B. Salwen, *An integrated approach to communication theory and research*. Routledge, 2014.

[24] S. Y. Rieh, "Encyclopedia of Library and Information Sciences," 2010.

[25] B. J. Fogg and H. Tseng, "The elements of computer credibility," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 1999, pp. 80–87.

[26] B. J. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, J. Shon, P. Swani, and others, "What makes Web sites credible?: a report on a large quantitative study," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2001, pp. 61–68.

[27] B. J. Fogg, "Persuasive technology: using computers to change what we think and do," *Ubiquity*, vol. 2002, no. December, p. 5, 2002.

[28] C. N. Wathen and J. Burkell, "Believe it or not: Factors influencing credibility on the Web," *J. Am. Soc. Inf. Sci. Technol.*, vol. 53, no. 2, pp. 134–144, 2002.

[29] Y. Yamamoto and K. Tanaka, "Enhancing credibility judgment of web search results," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 1235–1244.

[30] K. Kwon, J. Cho, and Y. Park, "Multidimensional credibility model for neighbor selection in collaborative recommendation," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 7114–7122, 2009.

[31] S. Y. Rieh and N. J. Belkin, “Understanding judgment of information quality and cognitive authority in the WWW,” in *Proceedings of the 61st annual meeting of the american society for information science*, 1998, vol. 35, pp. 279–289.

[32] Y. Gil and D. Artz, “Towards content trust of web resources,” *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 5, no. 4, pp. 227–239, 2007.

[33] V. L. Rubin and E. D. Liddy, “Assessing Credibility of Weblogs.,” in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 187–190.

[34] M. Kąkol, M. Jankowski-Lorek, K. Abramczuk, A. Wierzbicki, and M. Catasta, “On the subjectivity and bias of web content credibility evaluations,” in *Proceedings of the 22nd international conference on World Wide Web companion*, 2013, pp. 1131–1136.

[35] B. J. Fogg, J. Marshall, T. Kameda, J. Solomon, A. Rangnekar, J. Boyd, and B. Brown, “Web credibility research: a method for online experiments and early study results,” in *CHI’01 extended abstracts on Human factors in computing systems*, 2001, pp. 295–296.

[36] B. Bian, “Research of Factors on Impacting Internet Information Credibility Based on Electronic Commerce Users Demands,” in *Communication Systems and Network Technologies (CSNT), 2012 International Conference on*, 2012, pp. 987–991.

[37] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, “Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy,” in *Proceedings of the 22nd international conference on World Wide Web companion*, 2013, pp. 729–736.

[38] F. Yang, Y. Liu, X. Yu, and M. Yang, “Automatic Detection of Rumor on Sina Weibo,” in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, New York, NY, USA, 2012, p. 13:1–13:7.

[39] M. Mendoza, B. Poblete, and C. Castillo, “Twitter Under Crisis: Can we trust what we RT?,” in *Proceedings of the first workshop on social media analytics*, 2010, pp. 71–79.

[40] K. R. Canini, B. Suh, and P. L. Pirolli, “Finding credible information sources in social networks based on content and social structure,” in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, 2011, pp. 1–8.

[41] M. Gupta, P. Zhao, and J. Han, “Evaluating Event Credibility on Twitter.,” in *SDM*, 2012, pp. 153–164.

[42] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, “Truthy: Mapping the Spread of Astroturf in Microblog Streams,” in *Proceedings of the 20th International Conference Companion on World Wide Web*, New York, NY, USA, 2011, pp. 249–252.

[43] K. R. McKelvey and F. Menczer, “Truthy: Enabling the study of online social networks,” in *Proceedings of the 2013 conference on Computer supported cooperative work companion*, 2013, pp. 23–26.

[44] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, “Detecting and Tracking Political Abuse in Social Media.,” in *ICWSM*, 2011.

[45] S. Ravikumar, R. Balakrishnan, and S. Kambhampati, “Ranking tweets considering trust and relevance,” in *Proceedings of the Ninth International Workshop on Information Integration on the Web*, 2012, p. 4.

[46] G. Yin, F. Jiang, S. Cheng, X. Li, and X. He, “AUTrust: A Practical Trust Measurement for Adjacent Users in Social Networks,” in *Cloud and Green Computing (CGC), 2012 Second International Conference on*, 2012, pp. 360–367.

[47] B. Kang, S. Sikdar, T. Hollerer, J. ODonovan, and S. Adali, “Deconstructing Information Credibility on Twitter,” in *22nd International World Wide Web Conference (WWW)*, 2013.

[48] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[49] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[50] K. R. McKelvey and F. Menczer, “Truthy: Enabling the study of online social networks,” in *Proceedings of the 2013 conference on Computer supported cooperative work companion*, 2013, pp. 23–26.

[51] A. Pal and S. Counts, “Identifying topical authorities in microblogs,” in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 45–54.

[52] A. Pal and S. Counts, “What’s in a@ name? How Name Value Biases Judgment of Microblog Authors.,” in *ICWSM*, 2011.

[53] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, “Tweeting is believing?: understanding microblog credibility perceptions,” in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 2012, pp. 441–450.

[54] J. Yang, S. Counts, M. R. Morris, and A. Hoff, “Microblog credibility perceptions: Comparing the usa and china,” in *Proceedings of the 2013 conference on Computer supported cooperative work*, 2013, pp. 575–586.

[55] D. Westerman, P. R. Spence, and B. Van Der Heide, “A social network as information: The effect of system generated reports of connectedness on credibility on Twitter,” *Comput. Hum. Behav.*, vol. 28, no. 1, pp. 199–206, 2012.

[56] D. Hansen, B. Shneiderman, and M. A. Smith, *Analyzing social media networks with NodeXL: Insights from a connected world*. Morgan Kaufmann, 2010.

[57] S. Greengard, “Following the crowd,” *Commun. ACM*, vol. 54, no. 2, pp. 20–22, 2011.

[58] G. Demartini, “Hybrid human–machine information systems: Challenges and opportunities,” *Comput. Netw.*, 2015.

[59] J. Surowiecki, “The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations little,” *Brown ISBN 0-316-86173-1*, 2004.

[60] R. Snow, B. O’Connor, D. Jurafsky, and A. Y. Ng, “Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks,” in *Proceedings of the conference on empirical methods in natural language processing*, 2008, pp. 254–263.

[61] S. Sikdar, B. Kang, J. O’Donovan, T. Hollerer, and S. Adah, “Understanding information credibility on twitter,” in *Social Computing (SocialCom), 2013 International Conference on*, 2013, pp. 19–24.

[62] A. J. Flanagin and M. J. Metzger, “The perceived credibility of personal Web page information as influenced by the sex of the source,” *Comput. Hum. Behav.*, vol. 19, no. 6, pp. 683–701, 2003.

[63] L. R. Goldberg, *The International Personality Item Pool (IPIP)*. Internet site: http://ipip. ori. org/ipip, 1999.

[64] W. A. Scott, “Reliability of content analysis: The case of nominal scale coding.,” *Public Opin. Q.*, 1955.

[65] J. Cohen and others, “A coefficient of agreement for nominal scales,” *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[66] J. L. Fleiss, “Measuring nominal scale agreement among many raters.,” *Psychol. Bull.*, vol. 76, no. 5, p. 378, 1971.

[67] K. Krippendorff, “Reliability in content analysis,” *Hum. Commun. Res.*, vol. 30, no. 3, pp. 411–433, 2004.

[68] K. Krippendorff, *Content analysis: An introduction to its methodology*. Sage, 2012.

[69] A. Bermingham and A. F. Smeaton, “A study of inter-annotator agreement for opinion retrieval,” in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 784–785.

[70] R. Artstein and M. Poesio, “Inter-coder agreement for computational linguistics,” *Comput. Linguist.*, vol. 34, no. 4, pp. 555–596, 2008.

[71] K. De Swert, “Calculating inter-coder reliability in media content analysis using Krippendorff’s Alpha,” *Cent. Polit. Commun.*, pp. 1–15, 2012.

[72] K. Krippendorff, “Systematic and random disagreement and the reliability of nominal data,” *Commun. Methods Meas.*, vol. 2, no. 4, pp. 323–338, 2008.

[73] K. Krippendorff, “Computing Krippendorff’s alpha reliability,” *Dep. Pap. ASC*, p. 43, 2007.

[74] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *biometrics*, pp. 159–174, 1977.

[75] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, “Quality control in crowdsourcing systems: Issues and directions,” *IEEE Internet Comput.*, no. 2, pp. 76–81, 2013.

[76] P. Burnap and M. L. Williams, “Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making,” *Policy Internet*, 2015.

[77] N. Altrabsheh, M. Cocea, and S. Fallahkhair, “Sentiment analysis: towards a tool for analysing real-time students feedback,” in *Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on*, 2014, pp. 419–423.

[78] M. Allahbakhsh, A. Ignjatovic, B. Benatallah, E. B. Seyed-Mehdi-Reza Beheshti, E. Bertino, and N. Foo, “Collusion Detection in Online Rating Systems.,” in *APWeb*, 2013, pp. 196–207.

[79] A. F. Hayes and K. Krippendorff, “Answering the call for a standard reliability measure for coding data,” *Commun. Methods Meas.*, vol. 1, no. 1, pp. 77–89, 2007.

[80] M.-S. Shang, C.-H. Jin, T. Zhou, and Y.-C. Zhang, “Collaborative filtering based on multi-channel diffusion,” *Phys. Stat. Mech. Its Appl.*, vol. 388, no. 23, pp. 4867–4871, 2009.

[81] D. V. Cicchetti and S. A. Sparrow, “Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior.,” *Am. J. Ment. Defic.*, 1981.

[82] Y.-B. Zhou, T. Lei, and T. Zhou, “A robust ranking algorithm to spamming,” *EPL Europhys. Lett.*, vol. 94, no. 4, p. 48002, 2011.

[83] A. Ignjatovic, C. T. Lee, C. Kutay, H. Guo, and P. Compton, “Computing marks from multiple assessors using adaptive averaging,” in *International Conference on Engineering Education (ICEE)*, 2009.

[84] Y. Li, “Rating the Raters A new statistical method for evaluating,” in *2012 National Conference on Information Technology and Computer Science*, 2012.

[85] M. Allahbakhsh, A. Ignjatovic, B. Benatallah, S.-M.-R. Beheshti, N. Foo, and E. Bertino, “An analytic approach to people evaluation in crowdsourcing systems,” *ArXiv Prepr. ArXiv12113200*, 2012.

[86] M. Allahbakhsh and A. Ignjatovic, “An iterative method for calculating robust rating scores,” *Parallel Distrib. Syst. IEEE Trans. On*, vol. 26, no. 2, pp. 340–350, 2015.

[87] R. C. Mayer, J. H. Davis, and F. D. Schoorman, “An integrative model of organizational trust,” *Acad. Manage. Rev.*, vol. 20, no. 3, pp. 709–734, 1995.

[88] N. Hayeri, C. K. Chung, and J. W. Pennebaker, “The development of linguistic inquiry and word count (LIWC) for Arabic texts,” *Austin TX LIWC Net*, 2010.

[89] W. Weerkamp and M. de Rijke, “Credibility-inspired ranking for blog post retrieval,” *Inf. Retr.*, vol. 15, no. 3–4, pp. 243–277, 2012.

[90] A. Alarifi and M. Alsaleh, “Web Spam: A Study of the Page Language Effect on the Spam Detection Features,” in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, 2012, vol. 2, pp. 216–221.

[91] A. Abbasi, H. Chen, and A. Salem, “Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums,” *ACM Trans. Inf. Syst. TOIS*, vol. 26, no. 3, p. 12, 2008.

[92] A. Abbasi and H. Chen, “Applying authorship analysis to extremist-group web forum messages,” *Intell. Syst. IEEE*, vol. 20, no. 5, pp. 67–75, 2005.

[93] M. Al-Kabi, H. Wahsheh, I. Alsmadi, E. Al-Shawakfa, A. Wahbeh, and A. Al-Hmoud, “Content-based analysis to detect Arabic web spam,” *J. Inf. Sci.*, vol. 38, no. 3, pp. 284–296, 2012.

[94] A. Goweder and A. De Roeck, “Assessment of a significant Arabic corpus,” in *Arabic NLP Workshop at ACL/EACL*, 2001.

[95] A. A. AlMansour and C. S. Iliopoulos, “Using Arabic Microblogs Features in Determining Credibility,” in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 2015, pp. 1212–1219.

[96] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[97] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.

[98] H. Liu and R. Setiono, “Chi2: Feature selection and discretization of numeric attributes,” in *tai*, 1995, p. 388.

[99] M. A. Hall, “Correlation-based feature selection for machine learning,” The University of Waikato, 1999.

[100] K. Kira and L. A. Rendell, “A practical approach to feature selection,” in *Proceedings of the ninth international workshop on Machine learning*, 1992, pp. 249–256.

[101] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[102] G. H. John and P. Langley, “Estimating continuous distributions in Bayesian classifiers,” in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 1995, pp. 338–345.

[103] S. K. M. Yi, M. Steyvers, M. D. Lee, and M. J. Dry, “The wisdom of the crowd in combinatorial problems,” *Cogn. Sci.*, vol. 36, no. 3, pp. 452–470, 2012.

[104] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. R. Stat. Soc. Ser. B Methodol.*, pp. 1–38, 1977.

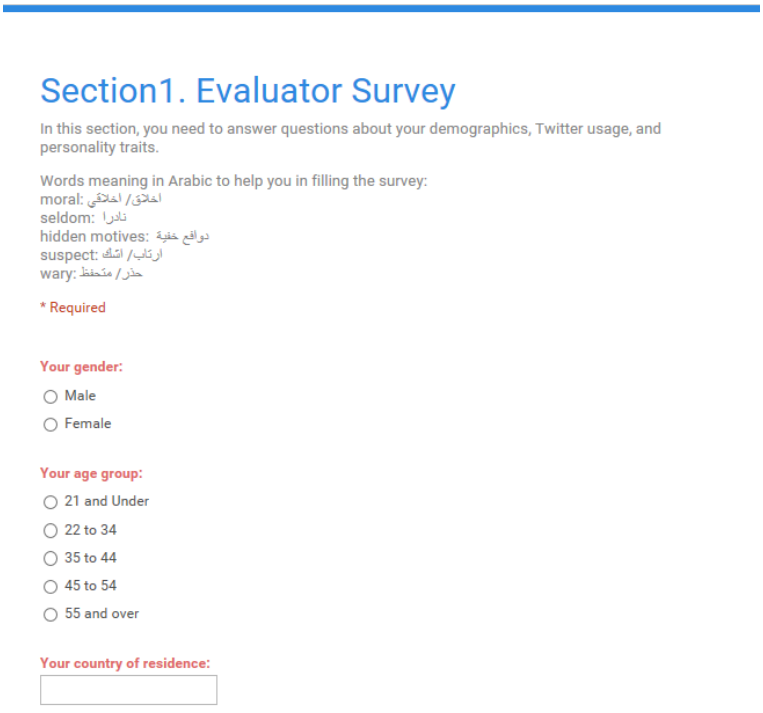
[105] R. Ennals, D. Byler, J. M. Agosta, and B. Rosario, “What is Disputed on the Web?,” in *Proceedings of the 4th workshop on Information credibility*, 2010, pp. 67–74.

Appendix A. **Survey Design**

This appendix contains snapshots of the survey designed to collect the tweet messages labelling. A complete listing for the survey used in this study can be found in this link: https://sites.google.com/site/arabiccredibilityproject/

1. Labeller Data Section

A complete listing for the labeller section (pre-labelling) questions used in the survey study is shown is below in the Figure A-1. In labeller section, participants answer user survey questions about their demographics, Twitter usage, and personal traits.



Section1. Evaluator Survey

In this section, you need to answer questions about your demographics, Twitter usage, and personality traits.

Words meaning in Arabic to help you in filling the survey:
moral: اخلاق/ اخلاقي
seldom: نادرا
hidden motives: دوافع خفية
suspect: ارتاب/ اشك
wary: حذر / متحفظ

* Required

Your gender:
- Male
- Female

Your age group:
- 21 and Under
- 22 to 34
- 35 to 44
- 45 to 54
- 55 and over

Your country of residence:

Your education level:

- ○ Did not graduate high school
- ○ High school graduate
- ○ Some college credit, no degree
- ○ College graduate
- ○ Postgraduate degree

Your major - If still in college/graduate/postgraduate:

Your Twitter username: *

Your Twitter display name:

- ○ Real name
- ○ Topical name (Name with topic nature such as politician_x, sport_expert)
- ○ Professional name (adding your profession such as doctor_x or lawyer_y)
- ○ Organizational (adding your organization such as KAU_x)
- ○ Nick name/ Any name

Your Twitter profile image:

- ○ Real photo
- ○ No image - Twitter profile image (Egg)
- ○ Image represent profession
- ○ Image represent organization
- ○ Any image

Which of the following biographical information you added into your Twitter profile Bio?

- ☐ Education
- ☐ Experience in a specific field
- ☐ Title or position of employment
- ☐ Contact information (e-mail or postal mail address, telephone number)
- ☐ Organizational authorship (corporate, governmental, or non-profit)
- ☐ None of the above

Is your Twitter account connected to your Facebook account?

- ○ Yes
- ○ No

You added your "webpage" on Twitter profile page:

- ○ Yes
- ○ No

You have a "verified" account:

- ○ Yes
- ○ No

Your followers count (Number of followers):

- ○ Number of followers <= 200
- ○ 200 < No. of followers <= 1000
- ○ 1000 < No. of followers <= 10000
- ○ 10000 < No. of followers

Your friends count (Number of following):

- ○ Number of following <= 200
- ○ 200 < No. of following <= 1000
- ○ 1000 < No. of following <= 10000
- ○ 10000 < No. of following

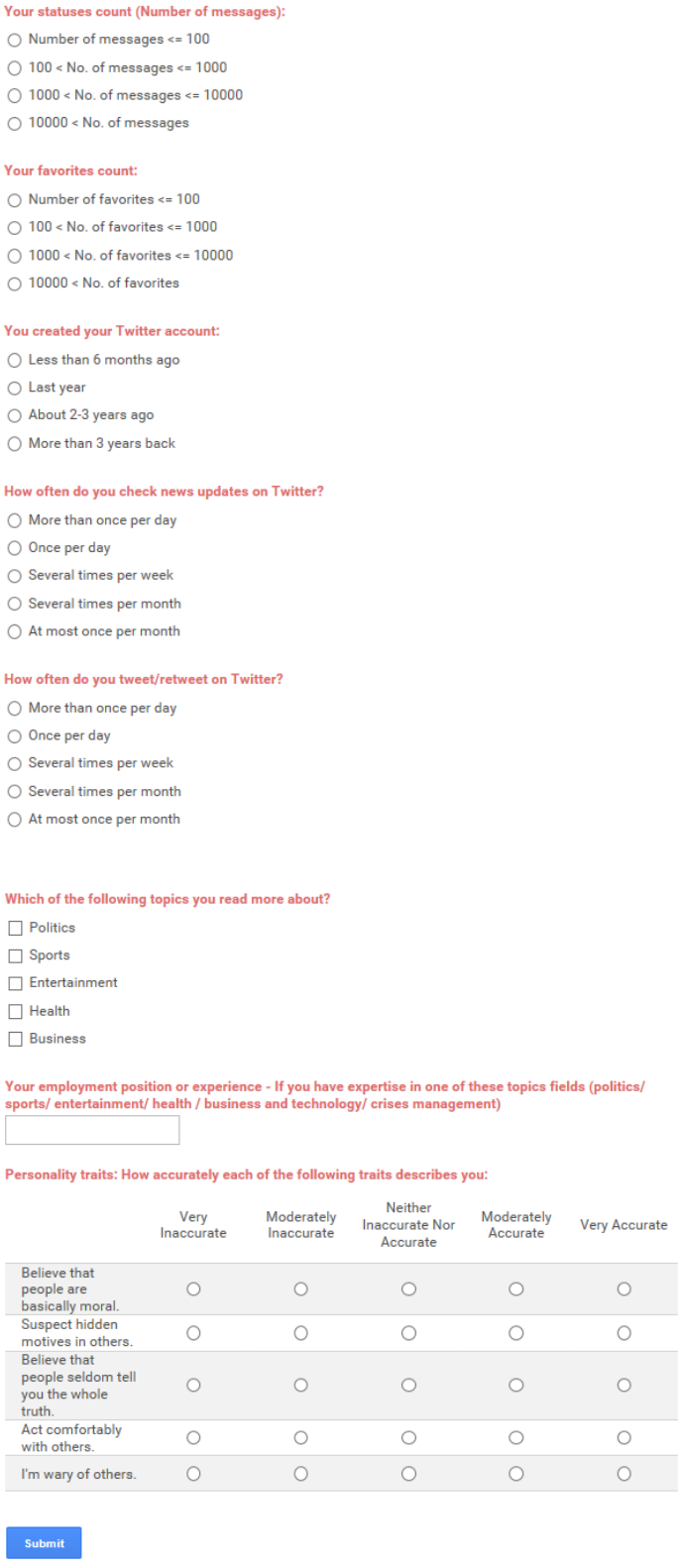


Figure A-1 The complete listing for the labeller section (pre-labelling)

2. Credibility Labelling Section

A partial listing for the labelling section using the topic: "Corona virus in Saudi Arabia / فيروس كورونا في السعودية" is shown below in Figure A-2 using Twitter presentation, and followed by the Text presentation in Figure A-3. Labelling section is the main part of the user study where

participants assign credibility degree to several tweets covering different topics based on the provided information (tweet content and author).

Section2a. Tweets' Credibility

In this section, you need to answer questions about specific topic familiarity and then assign credibility degree (1..5): 1= low credibility to 5= high credibility to several tweets based on the provided information (Tweet content and Author).

Topic: Corona virus in Saudi Arabia / فيروس كورونا في السعودية
Crises - Health - Domestic - April 2014

Words meaning in Arabic to help you in filling the survey:
unfamiliar: غير مألوف
cures/vaccines: علاج/ لقاحات
banned: حظر ومنع
outbreak: تفشي وانتشار
epidemic: وباء

* Required

Your Twitter username: *

How familiar you are with the "Corona virus in Saudi Arabia" news event?

	1 Unfamiliar with the news event	2	3	4	5 Know a lot about it
Topic Familiarity	○	○	○	○	○

How much would like to read more about the "Corona virus in Saudi Arabia" news event?

	1 Read Nothing	2	3	4	5 Read a lot about it
Topic Interest	○	○	○	○	○

How do you feel about the"Corona virus in Saudi Arabia" news event?

- ○ Mainly Negative
- ○ Mainly Positive
- ○ Neither/ Neutral
- ○ I Don't Care

Which is TRUE about the Corona virus in Saudi Arabia:

- ○ There are cures/vaccines being discovered for the virus
- ○ Travelers from Saudi Arabia are banned from entering certain countries
- ○ Health Minister was replaced as the number of infections and deaths increased.
- ○ Saudi Health Ministry says virus outbreak qualify it as an epidemic
- ○ I Don't know the answer

Tweet#1:

Figure A-2 A partial listing for topic#1 tweets labelling section using Twitter presentation

Credibility (1..5)	Tweet	Date	Tweeted by (User1)	Retweeted by/ Replied by (User2)	Type	Frequency	Link to User1	Link to User2	URLs in Tweet	Domains in Tweet
(1) Topic:Crises – Health - Domestic - April 2014 / فيروس كورونا في السعودية										
	RT @AlArabiya_KSA: ارتفاع الكورونا يخلق طوارىء مستشفى الملك فهد بجدة #مستشفى_الملك_فهد http://t.co/Xzr60qKcSm #كورونا_في_جدة	04/07/2014 23:02	alarabiya_ksa	om_talal	Retweet	40	User Inf.	User Inf.	http://www.alarabiya.net/ar/saudi-today/2014/04/08	alarabiya.net
	RT @almazroe50: اللهم سترك اللهم سترك عاجل إغلاق #مستشفى_الملك_فهد وذلك لانتشار فيروس كورونا وانفلونزا الخنازير (الي حاصل كارثة)وفقدان السيطرة على الوضع ،	04/07/2014 23:06	almazroe50	r07ee07m	Retweet	1	User Inf.	User Inf.		
	RT @k0a0kld2: شاهد فديو لمكتشف فيروس #مستشفى_الملك_فهد http://t.co/xK32av2weOجده	04/07/2014 23:09	k0a0kld2	r07ee07m	Retweet	1	User Inf.	User Inf.	http://www.youtube.com/watch?v=tYS4mwqZR5s&fe	youtube.com
	RT @smart_rn: عاجل من سبق : الان اغلاق مستشفي الملك فهد العام بجده. http://t.co/hTPtlkK8VC #مستشفى_الملك_فهد #وزارة_الصحة	04/07/2014 23:13	smart_rn	fahd_amer	Retweet	1	User Inf.	User Inf.	http://sabq.org/wsWfde	sabq.org
	RT @DR3lo: الاشتباه بإصابة ٣ أطباء وممرضة بـ " #كورونا " في مستشفى_الملك_فهد بـ #جدة# http://t.co/U7HYosGCna	04/08/2014 00:32	dr3lo	gogo1111688	Retweet	1	User Inf.	User Inf.	http://twasul.info/48634/	twasul.info
	#مستشفى_الملك_فهد بـ #جده									

Figure A-3 A partial listing for topic#1 tweets labelling section using Text presentation

3. Credibility Indicators Section

A complete listing for the credibility indicators (post-labelling) questions used in the survey study is shown below in Figure A-4. In this section, participants answer questions to indicate the importance of different features on assessing the information credibility.

Section3.Twitter Credibility Indicators Survey

In this section, you need to answer questions to indicate the importance of different features on assessing the information credibility.

Words meaning in Arabic to help you in filling the survey:
cues: اشارات
phrases: جمل/عبارات
trigger: محفز / مسبب

* Required

Your Twitter username: *

Which of the "Message Content" information you checked while deciding on credibility of message?

- [ ] Content
- [ ] Date
- [ ] Tweeted by (Author)
- [ ] Retweeted by/ Replied by (Other users)
- [ ] Retweet Count
- [ ] Favorite Count
- [ ] Users who retweeted / replied Information
- [ ] URLs in Tweet
- [ ] Domains in Tweet
- [ ] Hashtags in Tweet
- [ ] Twitter Page for Tweet

Which of the "Author" information you checked while deciding on credibility of message?

- [ ] None
- [ ] Image
- [ ] Verified (small blue checkmark)
- [ ] No. of Following
- [ ] No. of Followers
- [ ] No. of Tweets
- [ ] No. of Favorites
- [ ] Description
- [ ] Location
- [ ] Web
- [ ] Time Zone
- [ ] Joined Twitter Date
- [ ] Twitter Page for User

Are there any cues or phrases which trigger the sense of "5: high credibility" in the "Message Content"?

Are there any cues or phrases which trigger the sense of "1: low credibility" in the "Message Content"?

Are there any cues or phrases which trigger the sense of "5: high credibility" in the "Message's Author"?

Are there any cues or phrases which trigger the sense of "1: low credibility" in the "Message's Author"?

Content Credibility Features:

Identify how much "Content Features" indicate information credibility on a 5-point Likert scale ("1= implies low credibility" to "5=implies high credibility")

	1 Low Credibility	2	3	4	5 High Credibility
Message with formal language (No grammar or spelling mistakes):	○	○	○	○	○
Message with question mark "?":	○	○	○	○	○
Message with exclamation mark "!":	○	○	○	○	○
Message with personal pronouns – first person (as I), second person (as you), or third person (as he, she, it):	○	○	○	○	○
Message with emoticons such as sad "(:" or happy emoticons":)" :	○	○	○	○	○
Message with swear/bad words:	○	○	○	○	○
Message with unique special characters such as "€, ™, ¥, ©, ë, ...":	○	○	○	○	○
Message with longer length – more characters and words:	○	○	○	○	○
Message with hashtags such as #saudi, #boston_explosion	○	○	○	○	○
Message with URL:	○	○	○	○	○
Message with more re-tweets:	○	○	○	○	○
Message with more mentions "@":	○	○	○	○	○
Message with more "favorite" by other users:	○	○	○	○	○
Message posted recently:	○	○	○	○	○
Message content is similar with many tweets:	○	○	○	○	○
Message with event "image" attached:	○	○	○	○	○

Author Credibility Features:

Identify how much "Author Features" indicate information credibility on a 5-point Likert scale ("1= implies low credibility" to "5=implies high credibility")

	1 Low Credibility	2	3	4	5 High Credibility
Author usedh his/her real name:	○	○	○	○	○
Author used topical name (name with topic nature such as politician_x, sport_expert):	○	○	○	○	○

Figure A-4 The complete listing for the credibility indicators section (post-labelling)

4. Topic Labelling Section

A partial listing for the topic labelling section using the topic: "Preventing the use of 50 names for new-borns in Saudi Arabia / منع استخدام ٥٠ اسما للمواليد في السعودية" is shown below in Figure A-5. In this optional labelling section, participants need to answer questions about specific topic familiarity and then indicate the credibility degree for the topic (not the tweets) based on reading group of tweets.

Extra Section. Topics' Credibility

In this section, you need to answer questions about specific topic familiarity and then indicate the credibility degree (1..5): 1= low credibility to 5= high credibility for the TOPIC based on reading tweets below.

TOPIC: حضر استخدام 50 اسما للمواليد في السعودية

* Required

Your Twitter username: *

How familiar you are with "حضر 50 اسما للمواليد" news event?

	1 Unfamiliar with the news event	2	3	4	5 Know a lot about it
Topic Familiarity	○	○	○	○	○

How much would like to read more about the ""حضر 50 اسما للمواليد" news event?

	1 Read Nothing	2	3	4	5 Read a lot about it
Topic Interest	○	○	○	○	○

How do you feel about "حضر 50 اسما للمواليد" news event?

- ○ Mainly Negative
- ○ Mainly Positive
- ○ Neither/ Neutral
- ○ I Don't Care

Sample of tweets for this topic:

Figure A-5 A partial listing for topic labelling section

Appendix B. **Labellers' Weighting Results**

This appendix contains the complete results generated by the labellers' weighting models proposed in chapter 4.

1. Similarity and consistency model

The computed weights for all labellers after applying both pairwise rating and average rating similarity for different similarity measures are displayed below. Table B-1 shows the labellers' weights using pairwise rating similarity measures while Table B-2 shows the labellers' weights using average rating similarity measures. Moreover, the labellers' weights after applying the iterative algorithm using similarity model is displayed in Table B-3.

Table B-1 Labellers' weights using pairwise rating similarity measures

Judges	Cosine pairwise rating similarity	normalized	Pearson pairwise rating similarity	normalized	Jaccard pairwise rating similarity	normalized
Labeller1	0.6342	0.8989	0.1713	0.2497	0.4063	0.7957
Labeller2	0.6194	0.8629	0.1906	0.3230	0.4022	0.7829
Labeller3	0.5822	0.7730	0.1859	0.3051	0.3955	0.7620
Labeller4	0.6353	0.9016	0.1879	0.3125	0.4206	0.8403
Labeller5	0.6090	0.8378	0.1618	0.2135	0.4031	0.7857
Labeller6	0.6217	0.8686	0.2179	0.4267	0.4068	0.7972
Labeller7	0.5776	0.7618	0.2711	0.6286	0.3894	0.7432
Labeller8	0.5543	0.7053	0.2824	0.6714	0.3615	0.6565
Labeller9	0.5522	0.7003	0.2622	0.5948	0.3553	0.6371
Labeller10	0.6497	0.9363	0.1275	0.0832	0.4248	0.8532
Labeller11	0.6175	0.8583	0.2605	0.5883	0.4265	0.8585
Labeller12	0.5302	0.6469	0.1688	0.2402	0.3252	0.5437
Labeller13	0.6315	0.8923	0.2610	0.5901	0.4305	0.8710
Labeller14	0.5505	0.6960	0.1133	0.0293	0.3793	0.7118
Labeller15	0.5658	0.7331	0.3689	1.0000	0.3758	0.7008
Labeller16	0.5788	0.7647	0.1569	0.1949	0.4146	0.8215
Labeller17	0.6461	0.9277	0.2390	0.5065	0.4422	0.9074
Labeller18	0.5450	0.6827	0.1068	0.0047	0.3778	0.7072
Labeller19	0.5514	0.6984	0.3295	0.8503	0.3851	0.7297
Labeller20	0.6356	0.9023	0.2094	0.3943	0.4303	0.8702
Labeller21	0.6061	0.8307	0.1700	0.2447	0.4231	0.8478
Labeller22	0.5598	0.7186	0.3385	0.8843	0.3641	0.6645

Labeller23	0.5648	0.7307	0.3352	0.8717	0.3859	0.7323
Labeller24	0.5699	0.7432	0.2736	0.6381	0.3484	0.6156
Labeller25	0.5541	0.7049	0.2336	0.4861	0.3869	0.7354
Labeller26	0.5887	0.7887	0.1565	0.1934	0.3521	0.6273
Labeller27	0.5899	0.7917	0.1349	0.1116	0.3638	0.6634
Labeller28	0.5432	0.6785	0.3125	0.7857	0.3761	0.7017
Labeller29	0.5477	0.6894	0.1493	0.1661	0.3107	0.4985
Labeller30	0.5925	0.7979	0.2255	0.4554	0.3838	0.7258
Labeller31	0.6760	1.0000	0.1843	0.2991	0.4720	1.0000
Labeller32	0.3397	0.1856	0.1055	0.0000	0.2304	0.2487
Labeller33	0.5434	0.6789	0.1961	0.3439	0.3672	0.6742
Labeller34	0.6572	0.9545	0.2124	0.4057	0.4481	0.9257
Labeller35	0.3091	0.1115	0.2546	0.5657	0.1936	0.1343
Labeller36	0.2631	0.0000	0.1658	0.2289	0.1504	0.0000
Labeller37	0.6213	0.8675	0.2838	0.6767	0.4194	0.8365

Table B-2 Labellers' weights using average rating similarity measures

Judges	**Cosine average rating similarity**	**normalized**	**Pearson Average rating similarity**	**normalized**	**Jaccard average rating similarity**	**normalized**	**ICC average rating similarity**	**normalized**
Labeller1	0.9573	0.9885	0.5079	0.6382	0.8569	0.9339	0.2800	0.4824
Labeller2	0.9641	1.0000	0.5439	0.6988	0.8653	0.9443	0.2850	0.4934
Labeller3	0.8979	0.8891	0.4098	0.4732	0.8143	0.8814	0.2600	0.4383
Labeller4	0.9637	0.9992	0.4078	0.4698	0.9105	1.0000	0.3100	0.5485
Labeller5	0.9487	0.9741	0.3273	0.3344	0.8868	0.9708	0.2440	0.4031
Labeller6	0.9563	0.9869	0.5720	0.7461	0.8792	0.9615	0.3620	0.6630
Labeller7	0.9422	0.9632	0.6776	0.9237	0.8643	0.9431	0.4160	0.7819
Labeller8	0.4592	0.1542	0.5139	0.6482	0.1893	0.1109	0.2890	0.5022
Labeller9	0.4950	0.2142	0.4888	0.6060	0.2981	0.2450	0.2950	0.5154
Labeller10	0.8679	0.8388	0.1972	0.1155	0.6043	0.6224	0.0610	0.0000
Labeller11	0.8809	0.8605	0.6278	0.8399	0.7870	0.8477	0.4240	0.7996
Labeller12	0.5145	0.2469	0.3630	0.3945	0.2415	0.1751	0.2210	0.3524
Labeller13	0.8699	0.8421	0.4513	0.5431	0.6759	0.7108	0.2060	0.3194
Labeller14	0.3753	0.0137	0.2357	0.1803	0.1920	0.1141	0.1130	0.1145
Labeller15	0.3838	0.0279	0.7230	1.0000	0.1334	0.0419	0.4950	0.9559
Labeller16	0.3930	0.0433	0.2206	0.1549	0.1844	0.1048	0.1170	0.1233
Labeller17	0.8879	0.8723	0.5218	0.6617	0.7942	0.8567	0.3560	0.6498
Labeller18	0.3696	0.0041	0.2107	0.1382	0.1812	0.1009	0.1140	0.1167
Labeller19	0.3726	0.0091	0.6415	0.8630	0.1566	0.0705	0.4340	0.8216
Labeller20	0.8611	0.8274	0.2878	0.2680	0.6258	0.6491	0.1100	0.1079
Labeller21	0.8784	0.8564	0.4533	0.5463	0.6898	0.7280	0.2110	0.3304

Labeller22	0.3779	0.0180	0.6573	0.8895	0.1258	0.0326	0.4060	0.7599
Labeller23	0.3820	0.0249	0.6972	0.9566	0.1423	0.0529	0.5150	1.0000
Labeller24	0.3842	0.0286	0.5272	0.6706	0.1117	0.0152	0.2480	0.4119
Labeller25	0.3778	0.0178	0.5290	0.6738	0.1824	0.1023	0.3070	0.5419
Labeller26	0.6342	0.4474	0.2418	0.1906	0.3076	0.2567	0.0890	0.0617
Labeller27	0.5377	0.2858	0.2737	0.2442	0.3625	0.3244	0.0750	0.0308
Labeller28	0.3672	0.0002	0.6084	0.8073	0.1559	0.0696	0.3910	0.7269
Labeller29	0.3700	0.0048	0.2998	0.2882	0.0994	0.0000	0.1230	0.1366
Labeller30	0.5247	0.2640	0.5039	0.6314	0.3256	0.2789	0.3050	0.5374
Labeller31	0.8656	0.8349	0.2229	0.1588	0.7250	0.7713	0.1410	0.1762
Labeller32	0.6580	0.4873	0.1285	0.0000	0.4522	0.4350	NA	NA
Labeller33	0.3671	0.0000	0.3931	0.4451	0.1402	0.0503	0.2970	0.5198
Labeller34	0.8992	0.8912	0.4372	0.5192	0.8091	0.8750	0.3000	0.5264
Labeller35	0.4715	0.1749	0.5200	0.6586	0.2807	0.2235	0.2140	0.3370
Labeller36	0.4274	0.1009	0.4862	0.6017	0.2122	0.1391	0.2470	0.4097
Labeller37	0.8900	0.8759	0.7132	0.9835	0.7834	0.8433	0.3960	0.7379

Table B-3 Labellers' weights after applying the iterative algorithm using similarity model

Judges	**Cosine similarity algorithm**	**Pearson similarity algorithm**	**Jaccard similarity algorithm**
Labeller1	0.9873	0.4510	0.7981
Labeller2	1.0000	0.2264	0.8011
Labeller3	0.8919	0.3085	0.8760
Labeller4	0.9969	0.4349	0.9046
Labeller5	0.9607	0.1987	0.8697
Labeller6	0.9856	0.3819	0.8434
Labeller7	0.9514	0.2125	0.8228
Labeller8	0.1446	0.6493	0.1539
Labeller9	0.1955	0.7661	0.3000
Labeller10	0.9447	0.4245	0.8936
Labeller11	0.9592	0.2962	0.9482
Labeller12	0.2786	0.0000	0.2571
Labeller13	0.9370	0.3889	0.9559
Labeller14	0.0221	0.4129	0.1707
Labeller15	0.0367	0.9774	0.0697
Labeller16	0.0574	0.4888	0.1647
Labeller17	0.9799	0.4322	0.9425
Labeller18	0.0041	0.4198	0.1501
Labeller19	0.0092	0.9546	0.1100
Labeller20	0.9247	0.3357	0.9040
Labeller21	0.9528	0.1296	0.9711
Labeller22	0.0269	0.9361	0.0549
Labeller23	0.0338	1.0000	0.0873
Labeller24	0.0346	0.8501	0.0250
Labeller25	0.0275	0.7541	0.1576

Labeller26	0.5169	0.1945	0.3883
Labeller27	0.3372	0.3141	0.4114
Labeller28	0.0000	0.9118	0.1079
Labeller29	0.0105	0.5807	0.0000
Labeller30	0.3127	0.6332	0.3883
Labeller31	0.9389	0.6919	1.0000
Labeller32	0.4942	0.4096	0.4844
Labeller33	0.0022	0.6719	0.0811
Labeller34	0.9990	0.4348	0.9495
Labeller35	0.1499	0.7733	0.2435
Labeller36	0.0629	0.5518	0.1448
Labeller37	0.9802	0.3180	0.8781

2. Accuracy model

The computed weights for all labellers after applying pairwise rating accuracy measure is listed in Table B-4. On the other hand, the computed labellers' weights after applying average rating accuracy coupled with their algorithms are partitioned in to the following two tables: Table B-5 and Table B-6.

Table B-4 Labellers' weights using pairwise rating accuracy measures

Judges	Pairwise rating accuracy	normalized
Labeller1	0.138431	0.4369699
Labeller2	0.169949	0.502587
Labeller3	0.234133	0.6362069
Labeller4	0.265054	0.7005801
Labeller5	0.242379	0.6533735
Labeller6	0.226305	0.6199106
Labeller7	0.258621	0.6871866
Labeller8	0.373487	0.9263205
Labeller9	0.179775	0.5230427
Labeller10	0.22901	0.6255411
Labeller11	0.305476	0.784731
Labeller12	0.325523	0.8264677
Labeller13	0.274561	0.7203722
Labeller14	0.028816	0.2087694
Labeller15	0.408879	1
Labeller16	0.214953	0.5962778
Labeller17	0.268755	0.7082845
Labeller18	0.097522	0.3518041
Labeller19	0.334891	0.8459695
Labeller20	0.252547	0.6745419
Labeller21	0.279277	0.7301897
Labeller22	0.38785	0.9562229
Labeller23	0.405763	0.9935145
Labeller24	0.367601	0.9140672
Labeller25	0.205607	0.5768213
Labeller26	0.265501	0.7015099
Labeller27	-0.07146	0
Labeller28	0.30919	0.7924641
Labeller29	0.33908	0.8546913
Labeller30	0.22333	0.6137169
Labeller31	0.25727	0.6843749
Labeller32	0.323179	0.8215866

Labeller33	0.326938	0.8294135
Labeller34	0.248683	0.6664975
Labeller35	0.132768	0.4251816
Labeller36	0.145363	0.4514026
Labeller37	0.243385	0.655468

Table B-5 Labellers' weights using average rating accuracy measures (1)

Judges	Average absolute deviation		normalized	Normalized deviation	normalized	Normalized deviation algorithm	normalized
Labeller1	0.7373	0.2627	0.3321	0.0269	0.3321	0.0269	0.3322
Labeller2	0.6843	0.3157	0.4083	0.0270	0.4083	0.0270	0.4085
Labeller3	0.7591	0.2409	0.3008	0.0269	0.3008	0.0269	0.3007
Labeller4	0.4842	0.5158	0.6957	0.0272	0.6957	0.0272	0.6960
Labeller5	0.5637	0.4363	0.5815	0.0271	0.5815	0.0271	0.5816
Labeller6	0.6145	0.3855	0.5086	0.0271	0.5086	0.0271	0.5087
Labeller7	0.7089	0.2911	0.3729	0.0270	0.3729	0.0270	0.3731
Labeller8	0.6259	0.3741	0.4922	0.0271	0.4922	0.0271	0.4927
Labeller9	0.7184	0.2816	0.3593	0.0269	0.3593	0.0269	0.3595
Labeller10	0.7712	0.2288	0.2833	0.0269	0.2833	0.0269	0.2832
Labeller11	0.5690	0.4310	0.5739	0.0271	0.5739	0.0271	0.5740
Labeller12	0.6218	0.3782	0.4980	0.0271	0.4980	0.0271	0.4984
Labeller13	0.6988	0.3012	0.3875	0.0270	0.3875	0.0270	0.3873
Labeller14	0.9253	0.0747	0.0619	0.0267	0.0619	0.0267	0.0617
Labeller15	0.4702	0.5298	0.7159	0.0272	0.7159	0.0272	0.7163
Labeller16	0.4495	0.5505	0.7457	0.0273	0.7457	0.0273	0.7448
Labeller17	0.5559	0.4441	0.5928	0.0271	0.5928	0.0271	0.5931
Labeller18	0.7880	0.2120	0.2592	0.0269	0.2592	0.0269	0.2590
Labeller19	0.5945	0.4055	0.5374	0.0271	0.5374	0.0271	0.5380
Labeller20	0.7382	0.2618	0.3307	0.0269	0.3307	0.0269	0.3306
Labeller21	0.6560	0.3440	0.4489	0.0270	0.4489	0.0270	0.4488
Labeller22	0.5568	0.4432	0.5915	0.0271	0.5915	0.0271	0.5922
Labeller23	0.4639	0.5361	0.7250	0.0272	0.7250	0.0272	0.7253
Labeller24	0.5939	0.4061	0.5381	0.0271	0.5381	0.0271	0.5389
Labeller25	0.7055	0.2945	0.3778	0.0270	0.3778	0.0270	0.3776
Labeller26	0.7481	0.2519	0.3166	0.0269	0.3166	0.0269	0.3169
Labeller27	0.9684	0.0316	0.0000	0.0267	0.0000	0.0267	0.0000
Labeller28	0.6756	0.3244	0.4207	0.0270	0.4207	0.0270	0.4213
Labeller29	0.6494	0.3506	0.4584	0.0270	0.4584	0.0270	0.4592
Labeller30	0.6684	0.3316	0.4311	0.0270	0.4311	0.0270	0.4312
Labeller31	0.6039	0.3961	0.5238	0.0271	0.5238	0.0271	0.5236
Labeller32	0.2725	0.7275	1.0000	0.0275	1.0000	0.0275	1.0000
Labeller33	0.5460	0.4540	0.6070	0.0271	0.6070	0.0271	0.6074
Labeller34	0.4942	0.5058	0.6814	0.0272	0.6814	0.0272	0.6816
Labeller35	0.7277	0.2723	0.3460	0.0269	0.3460	0.0269	0.3462
Labeller36	0.8508	0.1492	0.1690	0.0268	0.1690	0.0268	0.1687
Labeller37	0.7144	0.2856	0.3649	0.0269	0.3649	0.0269	0.3652

Table B-6 Labellers' weights using average rating accuracy measures (2)

Judges	Variance	normalized	Variance by topic	normalized	SD1	normalized	SD2	normalized
Labeller1	0.0480	0.4559	0.0514	0.4373	0.5926	0.3594	0.5427	0.3226
Labeller2	0.0482	0.5373	0.0508	0.4494	0.6088	0.3885	0.5663	0.3621
Labeller3	0.0481	0.4880	0.0505	0.3527	0.5844	0.3446	0.5477	0.3310
Labeller4	0.0488	0.7884	0.0528	0.8113	0.7596	0.6601	0.7424	0.6569
Labeller5	0.0486	0.6844	0.0519	0.6150	0.7228	0.5938	0.6935	0.5750
Labeller6	0.0485	0.6462	0.0520	0.6400	0.6804	0.5174	0.6531	0.5073
Labeller7	0.0483	0.5617	0.0513	0.4874	0.6320	0.4302	0.6030	0.4235
Labeller8	0.0486	0.7052	0.0377	0.7983	0.7978	0.7287	0.7869	0.7313
Labeller9	0.0481	0.4909	0.0409	0.5264	0.6617	0.4838	0.6094	0.4342
Labeller10	0.0481	0.4847	0.0508	0.4057	0.5429	0.2698	0.5089	0.2660
Labeller11	0.0487	0.7241	0.0522	0.6912	0.7327	0.6116	0.7184	0.6167
Labeller12	0.0486	0.6805	0.0377	0.7085	0.7482	0.6395	0.7324	0.6401
Labeller13	0.0483	0.5770	0.0513	0.5017	0.6223	0.4129	0.5977	0.4147
Labeller14	0.0473	0.1691	0.0286	0.2132	0.4567	0.1148	0.3684	0.0308
Labeller15	0.0491	0.8709	0.0299	0.9966	0.9484	1.0000	0.9474	1.0000
Labeller16	0.0490	0.8543	0.0298	0.9781	0.8057	0.7430	0.7895	0.7357
Labeller17	0.0487	0.7163	0.0521	0.6464	0.7367	0.6187	0.7168	0.6140
Labeller18	0.0477	0.3341	0.0289	0.3975	0.6169	0.4030	0.5405	0.3190
Labeller19	0.0486	0.6991	0.0296	0.8048	0.7827	0.7015	0.7632	0.6916
Labeller20	0.0482	0.5246	0.0510	0.4560	0.5800	0.3366	0.5503	0.3353
Labeller21	0.0484	0.6257	0.0518	0.5720	0.6586	0.4783	0.6379	0.4820
Labeller22	0.0488	0.7805	0.0297	0.8957	0.8484	0.8198	0.8421	0.8238
Labeller23	0.0491	0.8739	0.0299	1.0000	0.8976	0.9084	0.8947	0.9119
Labeller24	0.0487	0.7302	0.0296	0.8396	0.8013	0.7350	0.7895	0.7357
Labeller25	0.0481	0.4911	0.0292	0.5727	0.7079	0.5669	0.6579	0.5154
Labeller26	0.0482	0.5235	0.0446	0.4394	0.5834	0.3428	0.5510	0.3365
Labeller27	0.0468	0.0000	0.0354	0.0000	0.4758	0.1490	0.3571	0.0120
Labeller28	0.0484	0.6067	0.0294	0.7018	0.7365	0.6184	0.7105	0.6035
Labeller29	0.0485	0.6512	0.0295	0.7514	0.7962	0.7260	0.7778	0.7161
Labeller30	0.0482	0.5515	0.0371	0.5960	0.6900	0.5348	0.6479	0.4987
Labeller31	0.0485	0.6651	0.0517	0.6141	0.6618	0.4839	0.6379	0.4820
Labeller32	0.0494	1.0000	0.0573	0.8997	0.9013	0.9150	0.8929	0.9087
Labeller33	0.0487	0.7360	0.0296	0.8460	0.8046	0.7411	0.7838	0.7262
Labeller34	0.0488	0.7546	0.0527	0.7858	0.7695	0.6779	0.7457	0.6623
Labeller35	0.0480	0.4695	0.0479	0.3842	0.5085	0.2079	0.4510	0.1690
Labeller36	0.0479	0.4004	0.0510	0.2111	0.3930	0.0000	0.3500	0.0000
Labeller37	0.0482	0.5479	0.0514	0.5088	0.5994	0.3716	0.5690	0.3666

3. Agreement model

The computed weights for all labellers after applying both modified agreement percentage and alpha agreement is listed in Table B-7.

Table B-7 Labellers' weights using agreement model

Judges	Modified agreement percentage	normalized	Alpha agreement	normalized
Labeller1	0.3589	0.4148	-0.0545	0.5260
Labeller2	0.3119	0.2677	-0.0986	0.4309
Labeller3	0.4357	0.6550	0.0431	0.7365
Labeller4	0.3479	0.3804	-0.0156	0.6100
Labeller5	0.3516	0.3918	-0.0012	0.6410
Labeller6	0.3608	0.4207	0.0067	0.6580
Labeller7	0.4433	0.6788	0.1463	0.9590
Labeller8	0.5460	1.0000	0.0181	0.6825
Labeller9	0.3921	0.5187	0.0423	0.7349
Labeller10	0.4333	0.6475	-0.1833	0.2484
Labeller11	0.4288	0.6333	0.1154	0.8925
Labeller12	0.4951	0.8408	-0.0927	0.4438
Labeller13	0.4485	0.6951	-0.0373	0.5631
Labeller14	0.3194	0.2912	-0.1128	0.4004
Labeller15	0.5169	0.9088	0.1040	0.8679
Labeller16	0.2263	0.0000	-0.2571	0.0892
Labeller17	0.3825	0.4886	0.0515	0.7546
Labeller18	0.3303	0.3253	-0.1103	0.4057
Labeller19	0.4968	0.8461	0.1653	1.0000
Labeller20	0.4465	0.6888	-0.1053	0.4165
Labeller21	0.4496	0.6985	-0.1641	0.2897
Labeller22	0.5337	0.9616	0.0845	0.8258
Labeller23	0.5096	0.8862	0.1272	0.9179
Labeller24	0.5305	0.9515	-0.0670	0.4992
Labeller25	0.4133	0.5850	0.0863	0.8296
Labeller26	0.4682	0.7565	-0.2169	0.1758
Labeller27	0.2559	0.0926	-0.2985	0.0000
Labeller28	0.5088	0.8837	0.1570	0.9821
Labeller29	0.5321	0.9567	-0.1629	0.2924
Labeller30	0.4114	0.5789	0.0620	0.7772
Labeller31	0.3805	0.4822	-0.0001	0.6434
Labeller32	0.3043	0.2441	NA	NA
Labeller33	0.4629	0.7402	0.0332	0.7151
Labeller34	0.3591	0.4153	0.0207	0.6882
Labeller35	0.2898	0.1986	-0.1189	0.3873
Labeller36	0.3654	0.4352	0.0282	0.7044
Labeller37	0.4312	0.6410	0.1533	0.9741

4. Majority consensus model

The computed labellers' weights after applying different methods and algorithms based on majority consensus model is listed in Table B-8.

Table B-8 Labellers' weights using majority consensus model

Judges	Exact class matching	normalized	Class ratio	normalized	Normalized class ratio	normalized	Normalized class ratio algorithm	normalized
Labeller1	0.5327	0.5654	0.4056	0.5019	0.6159	0.5381	0.5993	0.5206
Labeller2	0.3724	0.3767	0.3683	0.3821	0.5630	0.4175	0.5338	0.3960
Labeller3	0.6231	0.6719	0.4498	0.6438	0.6813	0.6870	0.6993	0.7110
Labeller4	0.4394	0.4555	0.3886	0.4473	0.5942	0.4886	0.5630	0.4516
Labeller5	0.4573	0.4766	0.3804	0.4210	0.5848	0.4671	0.5648	0.4550
Labeller6	0.4796	0.5029	0.4098	0.5153	0.6222	0.5523	0.6023	0.5263
Labeller7	0.6834	0.7430	0.4647	0.6916	0.7028	0.7360	0.7092	0.7299
Labeller8	0.9016	1.0000	0.5608	1.0000	0.8186	1.0000	0.8511	1.0000
Labeller9	0.4844	0.5085	0.4272	0.5712	0.6284	0.5664	0.6128	0.5463
Labeller10	0.6450	0.6977	0.4445	0.6268	0.6842	0.6937	0.7197	0.7498
Labeller11	0.6667	0.7232	0.4484	0.6391	0.6910	0.7092	0.6893	0.6919
Labeller12	0.7746	0.8504	0.5062	0.8249	0.7520	0.8483	0.7780	0.8609
Labeller13	0.7184	0.7842	0.4576	0.6688	0.7038	0.7384	0.7362	0.7813
Labeller14	0.3421	0.3410	0.3402	0.2917	0.5046	0.2842	0.4730	0.2802
Labeller15	0.7895	0.8679	0.5304	0.9026	0.7748	0.9002	0.7886	0.8811
Labeller16	0.0526	0.0000	0.2493	0.0000	0.3799	0.0000	0.3258	0.0000
Labeller17	0.5376	0.5712	0.4124	0.5235	0.6365	0.5849	0.6174	0.5551
Labeller18	0.3514	0.3518	0.3500	0.3232	0.5189	0.3169	0.4829	0.2990
Labeller19	0.6842	0.7439	0.5116	0.8420	0.7493	0.8421	0.7563	0.8196
Labeller20	0.6805	0.7395	0.4537	0.6564	0.6971	0.7231	0.7308	0.7709
Labeller21	0.7299	0.7977	0.4605	0.6781	0.7086	0.7492	0.7370	0.7829
Labeller22	0.8421	0.9299	0.5474	0.9571	0.8008	0.9596	0.8281	0.9561
Labeller23	0.7105	0.7749	0.5243	0.8827	0.7640	0.8755	0.7688	0.8434
Labeller24	0.8684	0.9609	0.5443	0.9472	0.7971	0.9511	0.8308	0.9613
Labeller25	0.4474	0.4649	0.4307	0.5824	0.6300	0.5701	0.6095	0.5401
Labeller26	0.6939	0.7553	0.4719	0.7148	0.7112	0.7552	0.7502	0.8080
Labeller27	0.1714	0.1399	0.2825	0.1065	0.4304	0.1152	0.3816	0.1063
Labeller28	0.7105	0.7749	0.5230	0.8786	0.7662	0.8806	0.7794	0.8635
Labeller29	0.8889	0.9850	0.5460	0.9527	0.8013	0.9607	0.8368	0.9727
Labeller30	0.5211	0.5518	0.4363	0.6005	0.6510	0.6181	0.6338	0.5862
Labeller31	0.5575	0.5946	0.4065	0.5048	0.6294	0.5687	0.6390	0.5962
Labeller32	0.3214	0.3166	0.3412	0.2950	0.5445	0.3752	0.5017	0.3349
Labeller33	0.6216	0.6702	0.4789	0.7371	0.7014	0.7330	0.7055	0.7228
Labeller34	0.4624	0.4827	0.3835	0.4310	0.5971	0.4951	0.5647	0.4547
Labeller35	0.2157	0.1921	0.3239	0.2394	0.4965	0.2659	0.4527	0.2416

Labeller36	0.6000	0.6447	0.3947	0.4670	0.6302	0.5706	0.6505	0.6181
Labeller37	0.6322	0.6826	0.4496	0.6431	0.6905	0.7079	0.6875	0.6885

5. Propensity to trust model

The computed labellers' average deviations used for detecting labellers' propensity to trust is shown below in Table B-9.

Table B-9 Labellers' average deviations - propensity to trust model

Judges	PropTrust (average absolute deviation)	PropTrustsign (average deviation)	Perc Class {1} %	PropTrustsign + PropTrust	Rank
Labeller1	0.7832	0.4966	23.12%	1.2798	32
Labeller2	0.7274	0.5876	11.73%	1.3150	33
Labeller3	0.8091	-0.2577	66.83%	0.5514	16
Labeller4	0.5152	0.2625	14.14%	0.7776	20
Labeller5	0.6004	0.1951	23.62%	0.7955	22
Labeller6	0.6522	0.3715	21.43%	1.0237	28
Labeller7	0.7543	0.1960	44.22%	0.9503	26
Labeller8	0.6526	-0.4527	88.52%	0.1998	10
Labeller9	0.7525	0.3624	34.38%	1.1149	31
Labeller10	**0.8121**	**-0.7525**	**90.53%**	**0.0597**	**3**
Labeller11	0.5988	0.0173	41.95%	0.6161	18
Labeller12	0.6456	-0.4284	77.46%	0.2172	11
Labeller13	**0.7361**	**-0.6309**	**82.76%**	**0.1051**	**7**
Labeller14	0.9535	0.6019	26.32%	1.5554	36
Labeller15	0.4845	-0.3472	71.05%	0.1374	8
Labeller16	0.4631	0.4121	0.00%	0.8752	24
Labeller17	0.5845	0.2040	26.59%	0.7885	21
Labeller18	0.8120	0.5034	24.32%	1.3155	34
Labeller19	0.6126	-0.0489	60.53%	0.5637	17
Labeller20	**0.7773**	**-0.6966**	**86.39%**	**0.0807**	**4**
Labeller21	**0.6901**	**-0.6060**	**78.16%**	**0.0841**	**5**
Labeller22	0.5738	-0.4287	81.58%	0.1451	9
Labeller23	0.4781	-0.2388	63.16%	0.2392	13
Labeller24	**0.6121**	**-0.5642**	**89.47%**	**0.0478**	**1**
Labeller25	0.7270	0.3578	36.84%	1.0848	29
Labeller26	**0.7815**	**-0.6973**	**93.88%**	**0.0843**	**6**
Labeller27	1.0058	0.9875	4.29%	1.9934	37
Labeller28	0.6963	-0.0489	65.79%	0.6473	19
Labeller29	**0.6692**	**-0.6103**	**94.44%**	**0.0589**	**2**
Labeller30	0.6938	0.2852	36.62%	0.9791	27
Labeller31	0.6361	-0.4039	59.77%	0.2322	12
Labeller32	**NA**	**NA**	**94.74%**	**NA**	**NA**
Labeller33	0.2883	0.1493	12.77%	0.4376	15
Labeller34	0.5626	-0.1727	59.46%	0.3899	14
Labeller35	0.5183	0.3037	14.45%	0.8220	23
Labeller36	0.7647	0.7248	3.92%	1.4895	35

Labeller37	0.8961	-0.0022	52.50%	0.8939	25
Labeller38	NA	NA	63.16%	NA	NA
Labeller39	0.7520	0.3445	35.06%	1.0965	30

Appendix C. **Credibility Detection Results**

This appendix contains the complete results generated by employing the statistical and machine learning models to detect Arabic messages credibility.

1. Evaluating features using relative frequencies

Table C-1 below is shown the relative frequencies of features across three classes {1, 2, 3}. Followed by a Table C-2 that lists the highest similarity and agreement values between labellers for the following measures: Pearson similarity, Cosine similarity, Jaccard similarity, and Alpha agreement.

Table C-1 The relative frequencies of features across three classes {1, 2, 3}

	Class 3 NA%	Class 3 A%	Class 2 NA%	Class 2 A%	Class 1 NA%	Class 1 A%
topic	25.5%	74.5%	31.4%	68.6%	65.5%	34.5%
content_rank	64.7%	35.3%	54.3%	45.7%	48.7%	51.3%
content_RetweetNo	82.4%	17.6%	94.3%	5.7%	89.4%	10.6%
content_FavNo	74.5%	25.5%	85.7%	14.3%	85.8%	14.2%
content_HashNo	66.7%	33.3%	45.7%	54.3%	70.8%	29.2%
content_SpellNo	82.4%	17.6%	80.0%	20.0%	71.7%	28.3%
content_QmarkNo	98.0%	2.0%	91.4%	8.6%	99.1%	0.9%
content_ExcmarkNo	96.1%	3.9%	85.7%	14.3%	89.4%	10.6%
content_EmotiNo	98.0%	2.0%	100.0%	0.0%	95.6%	4.4%
content_SpecialchNo	98.0%	2.0%	97.1%	2.9%	99.1%	0.9%
content_CharNo	43.1%	56.9%	45.7%	54.3%	39.8%	60.2%
content_WordsNo	51.0%	49.0%	60.0%	40.0%	49.6%	50.4%
content_HasURL	52.9%	47.1%	48.6%	51.4%	67.3%	32.7%
content_HasImage	64.7%	35.3%	82.9%	17.1%	85.0%	15.0%
content_PronounsTNo	84.3%	15.7%	85.7%	14.3%	66.4%	33.6%
content_PronounsDNo	72.5%	27.5%	80.0%	20.0%	64.6%	35.4%
content_SQuest	100.0%	0.0%	91.4%	8.6%	99.1%	0.9%
content_HasLaugh	100.0%	0.0%	100.0%	0.0%	99.1%	0.9%
content_DialWNo	86.3%	13.7%	88.6%	11.4%	78.8%	21.2%
content_BadSwearNo	100.0%	0.0%	100.0%	0.0%	98.2%	1.8%
content_ReligiousWNo	96.1%	3.9%	100.0%	0.0%	96.5%	3.5%
content_AllDial	82.4%	17.6%	88.6%	11.4%	75.2%	24.8%
content_HasUrgnews	92.2%	7.8%	94.3%	5.7%	93.8%	6.2%
content_Formal	94.1%	5.9%	97.1%	2.9%	96.5%	3.5%
author_Verif	84.3%	15.7%	74.3%	25.7%	96.5%	3.5%
author_DefImage	100.0%	0.0%	97.1%	2.9%	99.1%	0.9%
author_FwngNo	98.0%	2.0%	100.0%	0.0%	92.0%	8.0%
author_FlrNo	76.5%	23.5%	74.3%	25.7%	92.0%	8.0%

author_LogFlrNo	51.0%	49.0%	48.6%	51.4%	60.2%	39.8%
author_RatioFwFl	88.2%	11.8%	80.0%	20.0%	68.1%	31.9%
author_TweetsNo	58.8%	41.2%	51.4%	48.6%	79.6%	20.4%
author_FavNo	78.4%	21.6%	68.6%	31.4%	83.2%	16.8%
author_RatioTweetFav	92.2%	7.8%	91.4%	8.6%	91.2%	8.8%
author_News	56.9%	43.1%	45.7%	54.3%	69.0%	31.0%
author_HasBio	5.9%	94.1%	5.7%	94.3%	9.7%	90.3%
author_Edu	100.0%	0.0%	100.0%	0.0%	97.3%	2.7%
author_Emp	80.4%	19.6%	94.3%	5.7%	77.0%	23.0%
author_Contact	92.2%	7.8%	97.1%	2.9%	95.6%	4.4%
author_AllInf	72.5%	27.5%	94.3%	5.7%	71.7%	28.3%
author_AllInf2	100.0%	0.0%	97.1%	2.9%	98.2%	1.8%
author_HasWeb	43.1%	56.9%	34.3%	65.7%	64.6%	35.4%
author_YearsNo	52.9%	47.1%	57.1%	42.9%	69.9%	30.1%
author_DescRelate	64.7%	35.3%	60.0%	40.0%	77.9%	22.1%
author_LocationRelate	64.7%	35.3%	85.7%	14.3%	81.4%	18.6%
author_AllRelate	54.9%	45.1%	57.1%	42.9%	64.6%	35.4%
author_HasSpecialch	96.1%	3.9%	97.1%	2.9%	94.7%	5.3%

2. Labellers' similarity and agreement compared to messages' features occurrences

Table C-2 below is shown the highest similarity and agreement values between labellers for the following measures: Pearson similarity, Cosine similarity, Jaccard similarity, and Alpha agreement.

Table C-2 The highest similarity and agreement values between labellers

Judges	**Pearson similarity**	**Judges**	**Cosine Similarity**	**Judges**	**Jaccard Similarity**	**Judges**	**Alpha Agreement**
Labeller1	0.4187	Labeller1	0.9353	Labeller1	0.8783	Labeller1	0.3907
Labeller2	0.4109	Labeller2	0.9397	Labeller2	0.8822	Labeller2	0.3928
Labeller3	0.4878	Labeller3	0.9052	Labeller3	0.8079	Labeller3	0.3291
Labeller4	0.5887	Labeller4	0.9249	Labeller4	0.8468	Labeller4	0.5874
Labeller5	0.5465	Labeller5	0.9101	Labeller5	0.8348	Labeller5	0.3146
Labeller6	0.4717	Labeller6	0.9397	Labeller6	0.8822	Labeller6	0.4449
Labeller7	0.6549	Labeller7	0.9016	Labeller7	0.8202	Labeller7	0.6531
Labeller8	0.6177	Labeller8	0.7900	Labeller8	0.6356	Labeller8	0.5703
Labeller9	0.4956	Labeller9	0.7630	Labeller9	0.5700	Labeller9	0.4929
Labeller10	0.6062	Labeller10	0.9091	Labeller10	0.8000	Labeller10	0.3521
Labeller11	0.5391	Labeller11	0.9059	Labeller11	0.8249	Labeller11	0.4246
Labeller12	**0.8051**	Labeller12	0.8988	Labeller12	0.6608	Labeller12	0.4565
Labeller13	0.6130	Labeller13	0.9094	Labeller13	0.8044	Labeller13	0.5029
Labeller14	0.4273	Labeller14	0.9296	Labeller14	0.8524	Labeller14	0.3739
Labeller15	0.6730	Labeller15	0.9232	Labeller15	0.8333	Labeller15	0.5342
Labeller16	**0.8051**	Labeller16	0.9391	Labeller16	0.8524	Labeller16	0.3238
Labeller17	0.5107	Labeller17	0.9270	Labeller17	0.8634	Labeller17	0.5047
Labeller18	0.3408	Labeller18	0.9175	Labeller18	0.8426	Labeller18	0.2210
Labeller19	0.6201	Labeller19	0.9232	Labeller19	0.8440	Labeller19	0.6082
Labeller20	0.6992	Labeller20	0.8962	Labeller20	0.8000	**Labeller20**	**0.6584**

Labeller21	0.3123	Labeller21	0.9018	Labeller21	0.7955	Labeller21	0.4273
Labeller22	0.7259	**Labeller22**	**0.9562**	**Labeller22**	**0.8971**	Labeller22	0.5703
Labeller23	0.5391	Labeller23	0.9269	Labeller23	0.8370	Labeller23	0.4246
Labeller24	0.7259	**Labeller24**	**0.9562**	**Labeller24**	**0.8971**	Labeller24	0.5536
Labeller25	0.5655	Labeller25	0.9052	Labeller25	0.8238	Labeller25	0.5635
Labeller26	0.6992	Labeller26	0.7701	Labeller26	0.5642	**Labeller26**	**0.6584**
Labeller27	0.5040	Labeller27	0.8988	Labeller27	0.7720	Labeller27	0.3246
Labeller28	0.6201	Labeller28	0.9155	Labeller28	0.8440	Labeller28	0.6082
Labeller29	0.6062	Labeller29	0.9143	Labeller29	0.7895	Labeller29	0.3773
Labeller30	0.4758	Labeller30	0.8929	Labeller30	0.7720	Labeller30	0.4641
Labeller31	0.5636	Labeller31	0.9028	Labeller31	0.8044	Labeller31	0.5229
Labeller32	NA	Labeller32	NA	Labeller32	NA	Labeller32	0.3140
Labeller33	0.3117	Labeller33	0.7099	Labeller33	0.4978	Labeller33	0.4641
Labeller34	0.4244	Labeller34	0.9079	Labeller34	0.8073	Labeller34	0.4160
Labeller35	0.5887	Labeller35	0.9270	Labeller35	0.8634	Labeller35	0.5874
Labeller36	0.5040	Labeller36	0.6075	Labeller36	0.3916	Labeller36	0.3246
Labeller37	0.3646	Labeller37	0.5706	Labeller37	0.3922	Labeller37	0.3427
Labeller38	NA	Labeller38	NA	Labeller38	NA	Labeller38	0.4389
Labeller39	0.6549	Labeller39	0.9128	Labeller39	0.8340	Labeller39	0.6531

Tables below are shown messages' features distributions across the most similar and agreed labellers as follows: 1) Table C-3 shows the tweet messages' features occurrences between the most similar labellers (labeller#12 and abeller#16.) using Pearson similarity. 2) Table C-4 shows the tweet messages' features distributions across the most similar labellers (labeller#22 and labeller#24) using both Cosine similarity and Jaccard similarity. 3) Finally, the tweet messages' features distributions for the most agreed labellers (labeller#20 and labeller#26) using alpha agreement is shown in Table C-5.

Table C-3 The distribution of features across labeller#12 and labeller#16

Features	**Distribution of features % based on 3 credibility classes for labeler#12**			**Distribution of features% based on 3 credibility classes for labeler#16**		
	Class 3	Class 2	Class 1	Class 3	Class 2	Class 1
content_RetweetNo	NA	100.00	2.78	66.67	14.29	NA
content_FavNo		50.00	100.00	33.33	17.14	
content_HashNo		0.00	2.78	0.00	28.57	
content_SpellNo		100.00	5.56	66.67	45.71	
content_QmarkNo		0.00	11.11	0.00	5.71	
content_ExcmarkNo		0.00	2.78	0.00	11.43	
content_EmotiNo		0.00	0.00	0.00	2.86	
content_SpecialchNo		100.00	100.00	100.00	100.00	
content_CharNo		0.00	0.00	0.00	62.86	
content_WordsNo		0.00	0.00	0.00	57.14	
content_HasURL		100.00	5.56	100.00	28.57	
content_HasImage		0.00	44.44	0.00	5.71	
content_PronounsTNo		0.00	44.44	0.00	45.71	
content_PronounsDNo		0.00	2.78	0.00	45.71	
content_SQuest		0.00	0.00	0.00	2.86	

content_HasLaugh		100.00	100.00	100.00	100.00	
content_DialWNo		0.00	0.00	0.00	17.14	
content_DadSwearNo		100.00	100.00	100.00	100.00	
content_ReligiousWNo		0.00	27.78	0.00	14.29	
content_AllDeli		0.00	5.56	0.00	28.57	
content_HasUrgnews		50.00	0.00	33.33	5.71	
content_Formal		100.00	100.00	100.00	100.00	
author_Verif		50.00	2.78	33.33	0.00	
author_DefImage		0.00	88.89	0.00	2.86	
author_FwngNo		0.00	47.22	33.33	17.14	
author_FlrNo		50.00	0.00	66.67	8.57	
author_LogFlrNo		100.00	2.78	100.00	31.43	
author_RatioFwFl		0.00	100.00	0.00	25.71	
author_TweetsNo		50.00	0.00	33.33	22.86	
author_FavNo		0.00	44.44	0.00	20.00	
author_RatioTweetFav		50.00	0.00	33.33	45.71	
author_News		50.00	97.22	66.67	17.14	
author_HasBio		100.00	2.78	100.00	97.14	
author_Edu		0.00	52.78	0.00	2.86	
author_Emp		50.00	8.33	33.33	54.29	
author_Contact		0.00	58.33	33.33	5.71	
author_AllInf1		50.00	5.56	66.67	57.14	
author_AllInf2		0.00	25.00	0.00	5.71	
author_HasWeb		50.00	100.00	66.67	22.86	
author_YearsNo		100.00	0.00	100.00	48.57	
author_DescRelate		50.00	16.67	33.33	45.71	
author_LocationRelate		0.00	55.56	0.00	17.14	
author_AllRelate		50.00	8.33	33.33	57.14	
author_HasSpecialch		0.00	8.33	0.00	8.57	

Table C-4 The distribution of features across labeller#22 and labeller#24

Features	**Distribution of features % based on 3 credibility classes for labeler#22**			**Distribution of features% based on 3 credibility classes for labeler#24**		
	Class 3	Class 2	Class 1	Class 3	Class 2	Class 1
content_RetweetNo	0.00	66.67	14.71	33.33	25.00	16.13
content_FavNo	0.00	0.00	11.11	0.00	50.00	16.13
content_HashNo	100.00	0.00	14.29	33.33	75.00	19.35
content_SpellNo	0.00	33.33	26.98	0.00	50.00	51.61
content_QmarkNo	0.00	33.33	1.59	33.33	0.00	3.23
content_ExcmarkNo	0.00	0.00	6.35	0.00	0.00	12.90
content_EmotiNo	0.00	0.00	1.59	0.00	0.00	3.23
content_SpecialchNo	100.00	100.00	53.97	100.00	100.00	100.00
content_CharNo	0.00	0.00	34.92	0.00	25.00	67.74
content_WordsNo	0.00	0.00	31.75	0.00	25.00	61.29
content_HasURL	0.00	100.00	15.87	66.67	50.00	29.03
content_HasImage	100.00	0.00	1.59	33.33	0.00	3.23
content_PronounsTNo	0.00	33.33	23.81	33.33	25.00	45.16

content_PronounsDNo	0.00	0.00	25.40	0.00	25.00	48.39
content_SQuest	0.00	0.00	1.59	0.00	0.00	3.23
content_HasLaugh	100.00	100.00	53.97	100.00	100.00	100.00
content_DialWNo	0.00	0.00	9.52	0.00	0.00	19.35
content_DadSwearNo	100.00	100.00	53.97	100.00	100.00	100.00
content_ReligiousWNo	0.00	0.00	7.94	0.00	25.00	12.90
content_AllDeli	0.00	0.00	15.87	0.00	25.00	29.03
content_HasUrgnews	0.00	33.33	3.17	0.00	0.00	9.68
content_Formal	100.00	100.00	53.97	100.00	100.00	100.00
author_Verif	0.00	0.00	1.59	0.00	25.00	0.00
author_DefImage	0.00	0.00	1.59	0.00	0.00	3.23
author_FwngNo	100.00	0.00	9.52	33.33	25.00	16.13
author_FlrNo	0.00	0.00	7.94	0.00	50.00	9.68
author_LogFlrNo	100.00	100.00	15.87	100.00	50.00	29.03
author_RatioFwFl	100.00	0.00	12.70	33.33	25.00	22.58
author_TweetsNo	100.00	33.33	11.11	66.67	50.00	16.13
author_FavNo	100.00	33.33	7.94	66.67	0.00	16.13
author_RatioTweetFav	0.00	33.33	25.40	0.00	50.00	48.39
author_News	0.00	0.00	12.70	0.00	50.00	19.35
author_HasBio	100.00	100.00	52.38	100.00	100.00	96.77
author_Edu	0.00	0.00	1.59	0.00	0.00	3.23
author_Emp	0.00	66.67	28.57	33.33	50.00	54.84
author_Contact	0.00	0.00	4.76	0.00	0.00	9.68
author_AllInf1	0.00	66.67	31.75	33.33	50.00	61.29
author_AllInf2	0.00	0.00	3.17	0.00	0.00	6.45
author_HasWeb	0.00	33.33	14.29	33.33	25.00	25.81
author_YearsNo	100.00	100.00	25.40	100.00	50.00	48.39
author_DescRelate	0.00	66.67	23.81	33.33	50.00	45.16
author_LocationRelate	0.00	33.33	7.94	33.33	50.00	9.68
author_AllRelate	0.00	66.67	30.16	33.33	75.00	54.84
author_HasSpecialch	0.00	0.00	4.76	0.00	25.00	6.45

Table C-5 The distribution of features across labeller#20 and labeller#26

Features	**Distribution of features % based on 3 credibility classes for labeler#20**			**Distribution of features% based on 3 credibility classes for labeler#26**		
	Class 3	Class 2	Class 1	Class 3	Class 2	Class 1
content_RetweetNo	100.00	55.56	15.48	75.00	100.00	17.58
content_FavNo	100.00	55.56	13.10	75.00	100.00	15.56
content_HashNo	0.00	66.67	48.81	25.00	0.00	51.11
content_SpellNo	33.33	33.33	28.57	25.00	50.00	28.89
content_QmarkNo	0.00	11.11	3.57	25.00	0.00	3.33
content_ExcmarkNo	0.00	0.00	10.71	0.00	0.00	10.00
content_EmotiNo	0.00	0.00	1.19	0.00	0.00	1.11
content_SpecialchNo	0.00	0.00	1.19	0.00	0.00	1.11
content_CharNo	33.33	44.44	58.33	0.00	50.00	58.89
content_WordsNo	0.00	33.33	40.48	0.00	0.00	41.11
content_HasURL	100.00	44.44	53.57	100.00	100.00	51.11

content_HasImage	0.00	11.11	9.52	0.00	0.00	10.00
content_PronounsTNo	0.00	44.44	29.76	50.00	0.00	30.00
content_PronounsDNo	0.00	55.56	30.95	25.00	0.00	33.33
content_SQuest	0.00	11.11	3.57	25.00	0.00	3.33
content_HasLaugh	100.00	100.00	100.00	100.00	100.00	100.00
content_DialWNo	0.00	22.22	16.67	0.00	0.00	17.78
content_DadSwearNo	100.00	100.00	100.00	100.00	100.00	100.00
content_ReligiousWNo	0.00	11.11	5.95	0.00	0.00	6.67
content_AllDeli	0.00	33.33	21.43	0.00	0.00	23.33
content_HasUrgnews	33.33	0.00	5.95	25.00	0.00	5.56
content_Formal	0.00	0.00	3.57	0.00	0.00	3.33
author_Verif	100.00	44.44	2.38	100.00	100.00	3.33
author_DefImage	0.00	0.00	1.19	0.00	0.00	1.11
author_FwngNo	0.00	11.11	13.10	0.00	0.00	13.33
author_FlrNo	100.00	44.44	3.57	100.00	100.00	4.44
author_LogFlrNo	100.00	88.89	32.14	100.00	100.00	35.56
author_RatioFwFl	0.00	11.11	33.33	0.00	0.00	32.22
author_TweetsNo	66.67	33.33	25.00	75.00	50.00	24.44
author_FavNo	0.00	44.44	17.86	0.00	0.00	21.11
author_RatioTweetFav	66.67	33.33	4.76	75.00	50.00	5.56
author_News	100.00	55.56	32.14	100.00	100.00	32.22
author_HasBio	100.00	100.00	94.05	100.00	100.00	94.44
author_Edu	0.00	0.00	1.19	0.00	0.00	1.11
author_Emp	0.00	11.11	32.14	0.00	0.00	31.11
author_Contact	0.00	0.00	5.95	0.00	0.00	5.56
author_AllInf1	0.00	11.11	35.71	0.00	0.00	34.44
author_AllInf2	0.00	0.00	3.57	0.00	0.00	3.33
author_HasWeb	100.00	66.67	45.24	100.00	100.00	45.56
author_YearsNo	66.67	66.67	27.38	100.00	50.00	28.89
author_DescRelate	0.00	33.33	35.71	25.00	0.00	35.56
author_LocationRelate	0.00	33.33	8.33	0.00	0.00	11.11
author_AllRelate	0.00	44.44	40.48	25.00	0.00	41.11
author_HasSpecialch	0.00	0.00	4.76	0.00	0.00	4.44

3. Classification Results Using Machine Learning Approach

- **Before applying the proposed model**: Table C-6 shows the classifier outputs before applying the proposed model using Maj_Class2 and Maj_Hi labelling which based on simple majority voting method.

Table C-6 Classifier outputs using Maj_Class2 and Maj_Hi labelling

Maj_Class2

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

topic <= 2
|   content_favno <= 10
|   |   author_tweetsno <= 356: 2 (3.0/1.0)
|   |   author_tweetsno > 356: 1 (56.0/5.0)
|   content_favno > 10: 2 (2.0/1.0)
topic > 2
|   author_verif <= 0
|   |   author_descrelate <= 0
|   |   |   content_hashno <= 4
|   |   |   |   author_specialch <= 0
|   |   |   |   |   content_formal <= 0
|   |   |   |   |   |   author_allinf1 <= 0
|   |   |   |   |   |   |   content_emotino <= 0
|   |   |   |   |   |   |   |   content_wordsno <= 6: 2 (3.0/1.0)
|   |   |   |   |   |   |   |   content_wordsno > 6
|   |   |   |   |   |   |   |   |   content_retweetno <= 1
|   |   |   |   |   |   |   |   |   |   content_hasURL <= 0
|   |   |   |   |   |   |   |   |   |   |   author_favno <= 150: 1 (12.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   author_favno > 150: 2 (5.0/2.0)
|   |   |   |   |   |   |   |   |   |   content_hasURL > 0
|   |   |   |   |   |   |   |   |   |   |   topic <= 7: 3 (9.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   topic > 7: 1 (4.0/2.0)
|   |   |   |   |   |   |   |   |   content_retweetno > 1
|   |   |   |   |   |   |   |   |   |   content_hasimage <= 0: 1 (17.0/2.0)
|   |   |   |   |   |   |   |   |   |   content_hasimage > 0
|   |   |   |   |   |   |   |   |   |   |   author_fwngno <= 135: 3 (3.0)
|   |   |   |   |   |   |   |   |   |   |   author_fwngno > 135: 1 (6.0)
|   |   |   |   |   |   |   content_emotino > 0: 1 (2.0/1.0)
|   |   |   |   |   |   author_allinf1 > 0
|   |   |   |   |   |   |   content_favno <= 17: 3 (4.0/1.0)
|   |   |   |   |   |   |   content_favno > 17: 1 (3.0)
|   |   |   |   |   content_formal > 0
|   |   |   |   |   |   content_delicate <= 0: 2 (4.0/1.0)
|   |   |   |   |   |   content_delicate > 0: 1 (2.0)
|   |   |   |   author_specialch > 0
|   |   |   |   |   author_fwngno <= 454: 1 (3.0)
|   |   |   |   |   author_fwngno > 454: 2 (2.0)
|   |   |   content_hashno > 4
|   |   |   |   topic <= 7: 3 (2.0)
|   |   |   |   topic > 7: 2 (2.0)
|   |   author_descrelate > 0
|   |   |   content_emotino <= 0
|   |   |   |   content_alldelicate <= 0
|   |   |   |   |   content_hashno <= 2
|   |   |   |   |   |   author_yearsntwt <= 3
|   |   |   |   |   |   |   author_news <= 0
|   |   |   |   |   |   |   |   author_emp <= 0
|   |   |   |   |   |   |   |   |   content_favno <= 1: 3 (5.0/2.0)
|   |   |   |   |   |   |   |   |   content_favno > 1: 1 (4.0/1.0)
|   |   |   |   |   |   |   |   author_emp > 0: 3 (2.0)
|   |   |   |   |   |   |   author_news > 0
|   |   |   |   |   |   |   |   content_favno <= 0: 2 (3.0)
|   |   |   |   |   |   |   |   content_favno > 0: 1 (3.0/1.0)
|   |   |   |   |   |   author_yearsntwt > 3: 3 (3.0)
|   |   |   |   |   content_hashno > 2: 2 (11.0)
|   |   |   |   content_alldelicate > 0: 3 (5.0/1.0)
|   |   |   content_emotino > 0: 1 (2.0)
|   author_verif > 0
|   |   author_favno <= 3: 3 (6.0/2.0)
|   |   author_favno > 3
|   |   |   content_wordsno <= 12: 3 (3.0/1.0)
|   |   |   content_wordsno > 12: 2 (8.0)

Number of Leaves  :     32

Size of the tree :      63


Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         102               51.2563 %
Incorrectly Classified Instances        97               48.7437 %
Kappa statistic                          0.1713
Mean absolute error                      0.3385
Root mean squared error                  0.5304
Relative absolute error                 84.4833 %
Root relative squared error            118.6085 %
Coverage of cases (0.95 level)          67.3367 %
Mean rel. region size (0.95 level)      55.4439 %
Total Number of Instances              199

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area  PRC Area  Class
                 0.713     0.429     0.664       0.713    0.688       0.287    0.658     0.625     1
                 0.360     0.161     0.429       0.360    0.391       0.211    0.614     0.374     2
                 0.171     0.215     0.171       0.171    0.171       -0.044   0.419     0.190     3
Weighted Avg.    0.513     0.317     0.503       0.513    0.507       0.200    0.598     0.472
```

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 77 11 20 |  a = 1
 18 18 14 |  b = 2
 21 13  7 |  c = 3
```

Maj_Hi

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

topic <= 2
|   content_favno <= 10
|   |   author_tweetsno <= 356: 2 (3.0/1.0)
|   |   author_tweetsno > 356: 1 (56.0/5.0)
|   content_favno > 10: 2 (2.0/1.0)
topic > 2
|   author_verif <= 0
|   |   topic <= 7
|   |   |   content_spellno <= 1
|   |   |   |   content_hasimage <= 0
|   |   |   |   |   content_hasURL <= 0
|   |   |   |   |   |   content_formal <= 0
|   |   |   |   |   |   |   content_rank <= 4
|   |   |   |   |   |   |   |   author_fwngno <= 14: 1 (2.0)
|   |   |   |   |   |   |   |   author_fwngno > 14: 3 (5.0/1.0)
|   |   |   |   |   |   |   content_rank > 4: 1 (25.0/1.0)
|   |   |   |   |   |   content_formal > 0: 1 (3.0/2.0)
|   |   |   |   |   content_hasURL > 0
|   |   |   |   |   |   content_spellno <= 0
|   |   |   |   |   |   |   content_delicate <= 0
|   |   |   |   |   |   |   |   content_favno <= 17
|   |   |   |   |   |   |   |   |   author_ratiofwfl <= 0.64: 3 (10.0)
|   |   |   |   |   |   |   |   |   author_ratiofwfl > 0.64: 1 (5.0/1.0)
|   |   |   |   |   |   |   |   content_favno > 17: 1 (2.0)
|   |   |   |   |   |   |   content_delicate > 0: 1 (3.0)
|   |   |   |   |   |   content_spellno > 0: 3 (2.0)
|   |   |   |   content_hasimage > 0
|   |   |   |   |   author_yearsntwt <= 0: 2 (2.0)
|   |   |   |   |   author_yearsntwt > 0
|   |   |   |   |   |   author_contact <= 0
|   |   |   |   |   |   |   topic <= 4
|   |   |   |   |   |   |   |   author_descrelate <= 0: 1 (3.0)
|   |   |   |   |   |   |   |   author_descrelate > 0: 2 (4.0/2.0)
|   |   |   |   |   |   |   topic > 4: 3 (20.0/5.0)
|   |   |   |   |   |   author_contact > 0: 1 (2.0)
|   |   |   content_spellno > 1
|   |   |   |   author_favno <= 138: 1 (2.0)
|   |   |   |   author_favno > 138: 2 (4.0/1.0)
|   |   topic > 7
|   |   |   author_allrelate <= 0
|   |   |   |   author_hasweb <= 0: 1 (6.0/1.0)
|   |   |   |   author_hasweb > 0
|   |   |   |   |   content_charno <= 104: 1 (2.0)
|   |   |   |   |   content_charno > 104: 2 (4.0)
|   |   |   author_allrelate > 0
|   |   |   |   content_hashno <= 2: 3 (6.0/1.0)
|   |   |   |   content_hashno > 2: 2 (9.0/1.0)
|   author_verif > 0
|   |   author_contact <= 0
|   |   |   author_yearsntwt <= 4: 2 (5.0/1.0)
|   |   |   author_yearsntwt > 4
|   |   |   |   content_qmarkno <= 0
|   |   |   |   |   content_hasURL <= 0: 2 (2.0)
|   |   |   |   |   content_hasURL > 0: 3 (5.0)
|   |   |   content_qmarkno > 0: 2 (2.0)
|   |   author_contact > 0: 3 (3.0)

Number of Leaves  :     29

Size of the tree :      57

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         107               53.7688 %
Incorrectly Classified Instances        92               46.2312 %
Kappa statistic                          0.2192
Mean absolute error                      0.3321
Root mean squared error                  0.5199
Relative absolute error                 83.3385 %
Root relative squared error            116.5693 %
Coverage of cases (0.95 level)          68.8442 %
Mean rel. region size (0.95 level)      60.804  %
Total Number of Instances              199

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area  PRC Area  Class
                 0.694     0.374     0.688       0.694     0.691       0.321     0.649     0.612     1
                 0.278     0.123     0.333       0.278     0.303       0.167     0.597     0.258     2
                 0.400     0.264     0.367       0.400     0.383       0.133     0.531     0.326     3
Weighted Avg.    0.538     0.298     0.535       0.538     0.536       0.241     0.607     0.469
```

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 75  9 24 |  a = 1
 12 10 14 |  b = 2
 22 11 22 |  c = 3
```

- **After applying selective proposed weighting measures**: Table C-7 shows the classifier outputs using labelling obtained by similarity and accuracy weighting proposed measures.

Table C-7 Classifier outputs using similarity and accuracy measures labelling

Cosine Similarity Algorithm

```
===Classifier model (full training set) ===

J48 pruned tree
------------------

topic <= 7
|   author_verif <= 0
|   |   topic <= 2
|   |   |   content_hashno <= 7
|   |   |   |   content_wordsno <= 7
|   |   |   |   |   author_yearsntwt <= 2: 1 (4.0/1.0)
|   |   |   |   |   author_yearsntwt > 2: 3 (2.0)
|   |   |   |   content_wordsno > 7: 1 (49.0/3.0)
|   |   |   content_hashno > 7: 2 (2.0/1.0)
|   |   topic > 2
|   |   |   content_spellno <= 1
|   |   |   |   content_wordsno <= 6
|   |   |   |   |   topic <= 6: 2 (2.0)
|   |   |   |   |   topic > 6: 3 (3.0)
|   |   |   |   content_wordsno > 6
|   |   |   |   |   content_hasURL <= 0
|   |   |   |   |   |   author_hasbio <= 0: 1 (5.0)
|   |   |   |   |   |   author_hasbio > 0
|   |   |   |   |   |   |   author_edu <= 0
|   |   |   |   |   |   |   |   author_contact <= 0
|   |   |   |   |   |   |   |   |   author_emp <= 0
|   |   |   |   |   |   |   |   |   |   content_delicate <= 0
|   |   |   |   |   |   |   |   |   |   |   content_wordsno <= 17
|   |   |   |   |   |   |   |   |   |   |   |   content_rank <= 4: 2 (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   content_rank > 4: 1 (18.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   content_wordsno > 17
|   |   |   |   |   |   |   |   |   |   |   |   topic <= 4
|   |   |   |   |   |   |   |   |   |   |   |   |   content_rank <= 1: 3 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   content_rank > 1: 2 (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   topic > 4: 3 (7.0)
|   |   |   |   |   |   |   |   |   |   content_delicate > 0: 1 (11.0/1.0)
|   |   |   |   |   |   |   |   |   author_emp > 0
|   |   |   |   |   |   |   |   |   |   content_emotino <= 0: 3 (7.0/1.0)
|   |   |   |   |   |   |   |   |   |   content_emotino > 0: 1 (2.0)
|   |   |   |   |   |   |   |   author_contact > 0: 1 (2.0)
|   |   |   |   |   |   |   author_edu > 0: 1 (2.0)
|   |   |   |   |   content_hasURL > 0
|   |   |   |   |   |   content_delicate <= 0
|   |   |   |   |   |   |   author_ratiofavtweet <= 11.933333: 3 (7.0)
|   |   |   |   |   |   |   author_ratiofavtweet > 11.933333
|   |   |   |   |   |   |   |   author_descrelate <= 0
|   |   |   |   |   |   |   |   |   content_rank <= 5: 1 (4.0)
|   |   |   |   |   |   |   |   |   content_rank > 5: 3 (4.0)
|   |   |   |   |   |   |   |   author_descrelate > 0: 1 (2.0)
|   |   |   |   |   |   content_delicate > 0: 1 (4.0/1.0)
|   |   |   content_spellno > 1
|   |   |   |   author_favno <= 138: 1 (2.0)
|   |   |   |   author_favno > 138: 2 (4.0/1.0)
|   author_verif > 0
|   |   content_hashno <= 1
|   |   |   content_pronnoT <= 0
|   |   |   |   author_tweetsno <= 88521: 3 (2.0)
|   |   |   |   author_tweetsno > 88521: 1 (2.0)
|   |   |   content_pronnoT > 0: 1 (3.0)
|   |   content_hashno > 1
|   |   |   author_favno <= 3: 3 (4.0)
|   |   |   author_favno > 3: 2 (6.0/1.0)
topic > 7
|   content_formal <= 0
|   |   author_allrelate <= 0
|   |   |   author_hasweb <= 0: 1 (5.0)
|   |   |   author_hasweb > 0
|   |   |   |   author_tweetsno <= 61410: 2 (7.0/1.0)
|   |   |   |   author_tweetsno > 61410: 1 (2.0/1.0)
|   |   author_allrelate > 0
|   |   |   content_hashno <= 1: 3 (4.0/1.0)
|   |   |   content_hashno > 1: 2 (11.0/1.0)
|   content_formal > 0: 3 (2.0)

Number of Leaves  :     35

Size of the tree :      69


Time taken to build model: 0.02 seconds
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          111               55.7789 %
Incorrectly Classified Instances         88               44.2211 %
Kappa statistic                          0.2258
Mean absolute error                      0.3137
Root mean squared error                  0.505
Relative absolute error                 79.9665 %
Root relative squared error            114.1285 %
Coverage of cases (0.95 level)          71.3568 %
Mean rel. region size (0.95 level)      57.7889 %
Total Number of Instances              199

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area  PRC Area
Class
                 0.721     0.443     0.672       0.721     0.696       0.281     0.632      0.627
1
                 0.314     0.091     0.423       0.314     0.361       0.252     0.660      0.299
2
                 0.377     0.233     0.370       0.377     0.374       0.144     0.538      0.330
3
Weighted Avg.    0.558     0.325     0.548       0.558     0.551       0.239     0.612      0.490

=== Confusion Matrix ===

  a  b  c   <-- classified as
 80  8 23 |  a = 1
 13 11 11 |  b = 2
 26  7 20 |  c = 3
```

Jaccard Similarity Algorithm

```
== Classifier model (full training set) ===

J48 pruned tree
------------------

author_verif <= 0
|   topic <= 7
|   |   content_hasimage <= 0
|   |   |   author_contact <= 0
|   |   |   |   author_news <= 0
|   |   |   |   |   content_qmarkno <= 0: 1 (77.0/11.0)
|   |   |   |   |   content_qmarkno > 0: 2 (2.0/1.0)
|   |   |   |   author_news > 0
|   |   |   |   |   author_favno <= 167
|   |   |   |   |   |   author_tweetsno <= 36590: 1 (17.0/1.0)
|   |   |   |   |   |   author_tweetsno > 36590
|   |   |   |   |   |   |   content_retweetno <= 4: 3 (5.0)
|   |   |   |   |   |   |   content_retweetno > 4: 1 (4.0)
|   |   |   |   |   author_favno > 167: 3 (4.0/1.0)
|   |   |   author_contact > 0
|   |   |   |   content_rank <= 21: 2 (2.0/1.0)
|   |   |   |   content_rank > 21: 1 (2.0)
|   |   content_hasimage > 0
|   |   |   author_specialch <= 0
|   |   |   |   author_news <= 0
|   |   |   |   |   content_retweetno <= 37
|   |   |   |   |   |   author_locrelate <= 0
|   |   |   |   |   |   |   content_favno <= 2: 1 (5.0/2.0)
|   |   |   |   |   |   |   content_favno > 2: 3 (5.0)
|   |   |   |   |   |   author_locrelate > 0
|   |   |   |   |   |   |   content_hashno <= 2: 3 (8.0/1.0)
|   |   |   |   |   |   |   content_hashno > 2: 1 (2.0/1.0)
|   |   |   |   |   content_retweetno > 37: 1 (4.0)
|   |   |   |   author_news > 0
|   |   |   |   |   content_charno <= 106: 2 (3.0/1.0)
|   |   |   |   |   content_charno > 106: 1 (7.0)
|   |   |   author_specialch > 0: 1 (4.0/1.0)
|   topic > 7
|   |   author_allrelate <= 0
|   |   |   content_favno <= 1
|   |   |   |   author_hasweb <= 0: 1 (6.0)
|   |   |   |   author_hasweb > 0
|   |   |   |   |   content_rank <= 4: 2 (2.0)
|   |   |   |   |   content_rank > 4: 1 (2.0)
|   |   |   content_favno > 1: 2 (2.0)
|   |   author_allrelate > 0
|   |   |   content_hashno <= 1: 3 (3.0)
|   |   |   content_hashno > 1: 2 (12.0/3.0)
author_verif > 0
|   content_hashno <= 1
|   |   author_allrelate <= 0: 1 (6.0/1.0)
|   |   author_allrelate > 0: 2 (2.0/1.0)
|   content_hashno > 1
|   |   author_favno <= 3: 3 (4.0)
|   |   author_favno > 3: 2 (9.0/2.0)

Number of Leaves  :     26

Size of the tree :      51


Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          113               56.7839 %
Incorrectly Classified Instances         86               43.2161 %
```

```
Kappa statistic                          0.1862
Mean absolute error                      0.3067
Root mean squared error                  0.4933
Relative absolute error                 82.3644 %
Root relative squared error            114.5078 %
Coverage of cases (0.95 level)          73.8693 %
Mean rel. region size (0.95 level)      59.9665 %
Total Number of Instances               199

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area  PRC Area
Class
                 0.767     0.506     0.697       0.767    0.730       0.270    0.665     0.696
1
                 0.257     0.091     0.375       0.257    0.305       0.194    0.549     0.258
2
                 0.273     0.200     0.279       0.273    0.276       0.073    0.549     0.272
3
Weighted Avg.    0.568     0.366     0.548       0.568    0.555       0.213    0.619     0.525

=== Confusion Matrix ===

  a  b  c   <-- classified as
 92  9 19 |  a = 1
 14  9 12 |  b = 2
 26  6 12 |  c = 3
```

Variance Accuracy

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

topic <= 2
|   content_favno <= 10: 1 (59.0/4.0)
|   content_favno > 10: 2 (2.0/1.0)
topic > 2
|   author_verif <= 0
|   |   topic <= 7
|   |   |   content_hasimage <= 0
|   |   |   |   content_spellno <= 1
|   |   |   |   |   content_hasURL <= 0
|   |   |   |   |   |   content_formal <= 0
|   |   |   |   |   |   |   content_rank <= 4
|   |   |   |   |   |   |   |   author_fwngno <= 14: 1 (2.0)
|   |   |   |   |   |   |   |   author_fwngno > 14: 3 (5.0/1.0)
|   |   |   |   |   |   |   content_rank > 4: 1 (25.0/1.0)
|   |   |   |   |   |   content_formal > 0: 1 (3.0/2.0)
|   |   |   |   |   content_hasURL > 0
|   |   |   |   |   |   content_spellno <= 0
|   |   |   |   |   |   |   content_delicate <= 0
|   |   |   |   |   |   |   |   content_hashno <= 2
|   |   |   |   |   |   |   |   |   author_favno <= 0: 3 (5.0/1.0)
|   |   |   |   |   |   |   |   |   author_favno > 0: 1 (8.0/1.0)
|   |   |   |   |   |   |   |   content_hashno > 2: 3 (4.0)
|   |   |   |   |   |   |   content_delicate > 0: 1 (3.0)
|   |   |   |   |   |   content_spellno > 0: 3 (2.0)
|   |   |   |   content_spellno > 1
|   |   |   |   |   content_pronounsD <= 0: 1 (2.0)
|   |   |   |   |   content_pronounsD > 0: 2 (2.0)
|   |   |   content_hasimage > 0
|   |   |   |   author_contact <= 0
|   |   |   |   |   author_yearsntwt <= 0: 2 (2.0)
|   |   |   |   |   author_yearsntwt > 0
|   |   |   |   |   |   content_pronnoT <= 0
|   |   |   |   |   |   |   author_ratiofwfl <= 0: 2 (5.0/1.0)
|   |   |   |   |   |   |   author_ratiofwfl > 0
|   |   |   |   |   |   |   |   topic <= 4
|   |   |   |   |   |   |   |   |   content_favno <= 0: 2 (4.0/2.0)
|   |   |   |   |   |   |   |   |   content_favno > 0: 1 (3.0)
|   |   |   |   |   |   |   |   topic > 4
|   |   |   |   |   |   |   |   |   content_retweetno <= 24: 3 (12.0/1.0)
|   |   |   |   |   |   |   |   |   content_retweetno > 24: 1 (3.0)
|   |   |   |   |   |   content_pronnoT > 0: 3 (2.0)
|   |   |   |   author_contact > 0: 1 (2.0)
|   |   topic > 7
|   |   |   author_allrelate <= 0
|   |   |   |   content_favno <= 1
|   |   |   |   |   author_hasweb <= 0: 1 (6.0)
|   |   |   |   |   author_hasweb > 0
|   |   |   |   |   |   content_rank <= 4: 2 (2.0)
|   |   |   |   |   |   content_rank > 4: 1 (2.0)
|   |   |   |   content_favno > 1: 2 (2.0)
|   |   |   author_allrelate > 0
|   |   |   |   content_hashno <= 1: 3 (3.0)
|   |   |   |   content_hashno > 1: 2 (12.0/2.0)
|   author_verif > 0
|   |   author_favno <= 3: 3 (6.0/2.0)
|   |   author_favno > 3
|   |   |   content_wordsno <= 12: 3 (3.0/1.0)
|   |   |   content_wordsno > 12: 2 (8.0)

Number of Leaves  :     30

Size of the tree :      59


Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===
```

```
Correctly Classified Instances           121               60.804  %
Incorrectly Classified Instances         78                39.196  %
Kappa statistic                          0.2932
Mean absolute error                      0.2788
Root mean squared error                  0.4799
Relative absolute error                  71.902  %
Root relative squared error              109.118  %
Coverage of cases (0.95 level)           73.8693 %
Mean rel. region size (0.95 level)       55.4439 %
Total Number of Instances                199

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area  PRC Area
Class
                 0.816     0.412     0.727       0.816     0.769       0.417     0.716      0.695
1
                 0.333     0.113     0.419       0.333     0.371       0.242     0.606      0.315
2
                 0.326     0.163     0.375       0.326     0.349       0.171     0.578      0.310
3
Weighted Avg.    0.608     0.296     0.585       0.608     0.594       0.326     0.662      0.531

=== Confusion Matrix ===

  a  b  c   <-- classified as
 93  7 14 |  a = 1
 15 13 11 |  b = 2
 20 11 15 |  c = 3
```

Standard Deviation Accuracy

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

topic <= 2
|   content_favno <= 10: 1 (59.0/5.0)
|   content_favno > 10: 2 (2.0/1.0)
topic > 2
|   author_verif <= 0
|   |   author_descrelate <= 0
|   |   |   content_hashno <= 4
|   |   |   |   author_allinf1 <= 0
|   |   |   |   |   content_excmarkno <= 0
|   |   |   |   |   |   content_delicate <= 0
|   |   |   |   |   |   |   topic <= 5
|   |   |   |   |   |   |   |   author_yearsntwt <= 2
|   |   |   |   |   |   |   |   |   topic <= 4: 1 (7.0/2.0)
|   |   |   |   |   |   |   |   |   topic > 4: 3 (7.0/1.0)
|   |   |   |   |   |   |   |   author_yearsntwt > 2
|   |   |   |   |   |   |   |   |   content_pronounsD <= 0: 2 (5.0/2.0)
|   |   |   |   |   |   |   |   |   content_pronounsD > 0: 3 (3.0)
|   |   |   |   |   |   |   topic > 5
|   |   |   |   |   |   |   |   content_retweetno <= 5
|   |   |   |   |   |   |   |   |   content_favno <= 1
|   |   |   |   |   |   |   |   |   |   content_charno <= 106: 1 (7.0)
|   |   |   |   |   |   |   |   |   |   content_charno > 106
|   |   |   |   |   |   |   |   |   |   |   content_hashno <= 2: 2 (3.0)
|   |   |   |   |   |   |   |   |   |   |   content_hashno > 2: 1 (2.0)
|   |   |   |   |   |   |   |   |   content_favno > 1: 2 (2.0)
|   |   |   |   |   |   |   |   content_retweetno > 5
|   |   |   |   |   |   |   |   |   topic <= 6: 1 (5.0)
|   |   |   |   |   |   |   |   |   topic > 6
|   |   |   |   |   |   |   |   |   |   content_rank <= 9: 1 (3.0)
|   |   |   |   |   |   |   |   |   |   content_rank > 9: 3 (3.0)
|   |   |   |   |   |   |   content_delicate > 0: 1 (14.0/2.0)
|   |   |   |   |   content_excmarkno > 0
|   |   |   |   |   |   content_pronounsD <= 0
|   |   |   |   |   |   |   author_fwngno <= 613: 1 (3.0)
|   |   |   |   |   |   |   author_fwngno > 613: 2 (3.0)
|   |   |   |   |   |   content_pronounsD > 0: 1 (3.0)
|   |   |   |   author_allinf1 > 0
|   |   |   |   |   content_hasimage <= 0
|   |   |   |   |   |   content_rank <= 5: 1 (4.0/1.0)
|   |   |   |   |   |   content_rank > 5: 3 (2.0)
|   |   |   |   |   content_hasimage > 0: 1 (3.0)
|   |   |   content_hashno > 4
|   |   |   |   topic <= 7: 3 (2.0)
|   |   |   |   topic > 7: 2 (2.0)
|   |   author_descrelate > 0
|   |   |   content_emotino <= 0
|   |   |   |   author_emp <= 0
|   |   |   |   |   content_favno <= 0
|   |   |   |   |   |   content_hashno <= 2
|   |   |   |   |   |   |   content_retweetno <= 3
|   |   |   |   |   |   |   |   author_tweetsno <= 16712: 2 (3.0)
|   |   |   |   |   |   |   |   author_tweetsno > 16712: 3 (3.0/1.0)
|   |   |   |   |   |   |   content_retweetno > 3: 3 (4.0)
|   |   |   |   |   |   content_hashno > 2: 2 (7.0)
|   |   |   |   |   content_favno > 0
|   |   |   |   |   |   content_wordsno <= 15: 2 (5.0/1.0)
|   |   |   |   |   |   content_wordsno > 15: 1 (6.0/1.0)
|   |   |   |   author_emp > 0: 3 (8.0/1.0)
|   |   |   content_emotino > 0: 1 (2.0)
|   author_verif > 0
|   |   author_contact <= 0
|   |   |   author_yearsntwt <= 4: 2 (5.0/1.0)
|   |   |   author_yearsntwt > 4
|   |   |   |   content_qmarkno <= 0
|   |   |   |   |   content_hasURL <= 0: 2 (2.0)
|   |   |   |   |   content_hasURL > 0: 3 (5.0/1.0)
|   |   |   |   content_qmarkno > 0: 2 (2.0)
```

```
|   |   author_contact > 0: 3 (3.0)

Number of Leaves  :     35

Size of the tree :      69


Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         110               55.2764 %
Incorrectly Classified Instances       89                44.7236 %
Kappa statistic                        0.2541
Mean absolute error                    0.3222
Root mean squared error                0.5173
Relative absolute error                81.1353 %
Root relative squared error            116.1846 %
Coverage of cases (0.95 level)         66.8342 %
Mean rel. region size (0.95 level)     55.7789 %
Total Number of Instances              199

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area  PRC Area
Class
                 0.682     0.360     0.701       0.682    0.691       0.321    0.628     0.621
1
                 0.395     0.199     0.354       0.395    0.374       0.189    0.538     0.273
2
                 0.391     0.170     0.409       0.391    0.400       0.225    0.600     0.307
3
Weighted Avg.    0.553     0.281     0.559       0.553    0.555       0.271    0.602     0.473

=== Confusion Matrix ===

  a  b  c   <-- classified as
 75 18 17 |  a = 1
 17 17  9 |  b = 2
 15 13 18 |  c = 3
```

- **After applying the proposed weighting aggregation model**: Table C-8 shows the classifier outputs using labelled dataset constructed by combining labellers' weights from different measures (similarity, accuracy, agreement, and majority consensus).

Table C-8 Classifier outputs using weighting aggregation model labelling

Weighting aggregation model - All features

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

topic <= 2
|   content_favno <= 10: 1 (59.0/5.0)
|   content_favno > 10: 2 (2.0/1.0)
topic > 2
|   topic <= 7
|   |   content_spellno <= 1
|   |   |   author_verif <= 0
|   |   |   |   content_hasURL <= 0
|   |   |   |   |   content_hasimage <= 0
|   |   |   |   |   |   content_formal <= 0
|   |   |   |   |   |   |   content_rank <= 4
|   |   |   |   |   |   |   |   author_fwngno <= 14: 1 (2.0)
|   |   |   |   |   |   |   |   author_fwngno > 14: 3 (5.0/1.0)
|   |   |   |   |   |   |   content_rank > 4: 1 (25.0/1.0)
|   |   |   |   |   |   content_formal > 0: 1 (3.0/2.0)
|   |   |   |   |   content_hasimage > 0
|   |   |   |   |   |   author_contact <= 0
|   |   |   |   |   |   |   content_wordsno <= 10
|   |   |   |   |   |   |   |   author_yearsntwt <= 1: 2 (2.0)
|   |   |   |   |   |   |   |   author_yearsntwt > 1
|   |   |   |   |   |   |   |   |   author_tweetsno <= 16712: 3 (3.0)
|   |   |   |   |   |   |   |   |   author_tweetsno > 16712: 2 (2.0)
|   |   |   |   |   |   |   content_wordsno > 10
|   |   |   |   |   |   |   |   author_emp <= 0
|   |   |   |   |   |   |   |   |   content_wordsno <= 18
|   |   |   |   |   |   |   |   |   |   author_allrelate <= 0
|   |   |   |   |   |   |   |   |   |   |   topic <= 4: 1 (3.0)
|   |   |   |   |   |   |   |   |   |   |   topic > 4: 3 (2.0)
|   |   |   |   |   |   |   |   |   |   author_allrelate > 0: 1 (5.0)
|   |   |   |   |   |   |   |   |   content_wordsno > 18: 3 (8.0/2.0)
|   |   |   |   |   |   |   |   author_emp > 0: 3 (4.0)
|   |   |   |   |   |   author_contact > 0: 1 (2.0)
|   |   |   |   content_hasURL > 0
|   |   |   |   |   content_spellno <= 0
|   |   |   |   |   |   content_delicate <= 0
|   |   |   |   |   |   |   content_favno <= 17
|   |   |   |   |   |   |   |   author_ratiofwfl <= 0.64: 3 (10.0)
|   |   |   |   |   |   |   |   author_ratiofwfl > 0.64: 1 (5.0/1.0)
|   |   |   |   |   |   |   content_favno > 17: 1 (2.0)
|   |   |   |   |   |   content_delicate > 0: 1 (3.0)
|   |   |   |   |   content_spellno > 0: 3 (2.0)
```

```
|   |   |   author_verif > 0
|   |   |   |   content_charno <= 98: 1 (2.0)
|   |   |   |   content_charno > 98
|   |   |   |   |   author_ratiofavtweet <= 21996.5: 3 (7.0/1.0)
|   |   |   |   |   author_ratiofavtweet > 21996.5: 2 (3.0)
|   |   content_spellno > 1
|   |   |   author_favno <= 138: 1 (2.0)
|   |   |   author_favno > 138: 2 (5.0/1.0)
|   topic > 7
|   |   author_allrelate <= 0
|   |   |   author_hasweb <= 0: 1 (6.0)
|   |   |   author_hasweb > 0
|   |   |   |   author_tweetsno <= 61410: 2 (7.0/1.0)
|   |   |   |   author_tweetsno > 61410: 1 (2.0/1.0)
|   |   author_allrelate > 0
|   |   |   content_hashno <= 1: 3 (4.0/1.0)
|   |   |   content_hashno > 1: 2 (12.0/2.0)

Number of Leaves  :     30

Size of the tree :      59


Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         117               58.794  %
Incorrectly Classified Instances        82               41.206  %
Kappa statistic                          0.281
Mean absolute error                      0.301
Root mean squared error                  0.4968
Relative absolute error                 77.52   %
Root relative squared error            112.8718 %
Coverage of cases (0.95 level)          71.3568 %
Mean rel. region size (0.95 level)      59.6315 %
Total Number of Instances              199

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area  PRC Area  Class
                 0.743     0.384     0.718       0.743    0.730       0.362    0.689     0.670     1
                 0.314     0.122     0.355       0.314    0.333       0.202    0.528     0.232     2
                 0.431     0.196     0.431       0.431    0.431       0.235    0.556     0.349     3
Weighted Avg.    0.588     0.290     0.581       0.588    0.584       0.301    0.627     0.511

=== Confusion Matrix ===

  a  b  c   <-- classified as
 84 11 18 |  a = 1
 13 11 11 |  b = 2
 20  9 22 |  c = 3
```

Weighting aggregation model - Relief Algorithm feature set

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:                                  CRED-weka.filters.unsupervised.attribute.Remove-R47-49-
weka.filters.supervised.attribute.AttributeSelection-
Eweka.attributeSelection.ReliefFAttributeEval -M -1 -D 1 -K 10-Sweka.attributeSelection.Ranker
-T -1.7976931348623157E308 -N -1-weka.filters.unsupervised.attribute.Remove-R15-46
Instances:    199
Attributes:   15
              topic
              content_hasimage
              author_descrelate
              content_hasURL
              author_locrelate
              author_emp
              author_hasweb
              content_wordsno
              content_alldelicate
              content_hashno
              author_allinf1
              author_allrelate
              author_logflrno
              author_verif
              sum
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

topic <= 2: 1 (61.0/7.0)
topic > 2
|   topic <= 7
|   |   author_verif <= 0
|   |   |   content_hasimage <= 0
|   |   |   |   content_hasURL <= 0
|   |   |   |   |   author_emp <= 0: 1 (33.0/7.0)
|   |   |   |   |   author_emp > 0
|   |   |   |   |   |   author_logflrno <= 3.8308: 1 (2.0)
|   |   |   |   |   |   author_logflrno > 3.8308: 3 (2.0)
|   |   |   |   content_hasURL > 0
|   |   |   |   |   content_hashno <= 3
|   |   |   |   |   |   content_hashno <= 2
|   |   |   |   |   |   |   topic <= 5
```

```
|   |   |   |   |   |   |   |   content_alldelicate <= 0: 3 (11.0/3.0)
|   |   |   |   |   |   |   |   content_alldelicate > 0: 1 (3.0/1.0)
|   |   |   |   |   |   |   topic > 5: 1 (4.0)
|   |   |   |   |   |   content_hashno > 2: 3 (4.0)
|   |   |   |   |   content_hashno > 3: 1 (2.0/1.0)
|   |   |   content_hasimage > 0
|   |   |   |   content_wordsno <= 9
|   |   |   |   |   topic <= 6: 2 (4.0/1.0)
|   |   |   |   |   topic > 6: 3 (3.0/1.0)
|   |   |   |   content_wordsno > 9
|   |   |   |   |   topic <= 4: 1 (8.0/3.0)
|   |   |   |   |   topic > 4
|   |   |   |   |   |   topic <= 5: 3 (7.0)
|   |   |   |   |   |   topic > 5
|   |   |   |   |   |   |   topic <= 6: 1 (5.0)
|   |   |   |   |   |   |   topic > 6: 3 (6.0/1.0)
|   |   author_verif > 0
|   |   |   author_allinf1 <= 0
|   |   |   |   content_alldelicate <= 0
|   |   |   |   |   content_wordsno <= 14: 3 (4.0/1.0)
|   |   |   |   |   content_wordsno > 14: 2 (4.0/1.0)
|   |   |   |   content_alldelicate > 0: 2 (2.0)
|   |   |   author_allinf1 > 0: 3 (3.0)
|   topic > 7
|   |   author_allrelate <= 0
|   |   |   author_hasweb <= 0: 1 (6.0)
|   |   |   author_hasweb > 0: 2 (9.0/3.0)
|   |   author_allrelate > 0
|   |   |   content_hashno <= 1: 3 (4.0/1.0)
|   |   |   content_hashno > 1: 2 (12.0/2.0)

Number of Leaves  :     23

Size of the tree :      45


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         131               65.8291 %
Incorrectly Classified Instances        68               34.1709 %
Kappa statistic                          0.3891
Mean absolute error                      0.2749
Root mean squared error                  0.4355
Relative absolute error                 70.7865 %
Root relative squared error             98.9265 %
Coverage of cases (0.95 level)          82.9146 %
Mean rel. region size (0.95 level)      74.3719 %
Total Number of Instances              199

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area  PRC Area  Class
                 0.823     0.372     0.744       0.823    0.782       0.462    0.751     0.737     1
                 0.429     0.079     0.536       0.429    0.476       0.382    0.678     0.366     2
                 0.451     0.155     0.500       0.451    0.474       0.306    0.623     0.416     3
Weighted Avg.    0.658     0.265     0.645       0.658    0.649       0.408    0.705     0.590

=== Confusion Matrix ===

  a  b  c   <-- classified as
 93  6 14 |  a = 1
 11 15  9 |  b = 2
 21  7 23 |  c = 3
```

- **Classification results using different classification algorithms**: Table C-9 shows the classifier outputs for different settings and algorithms using two credibility classes {1, 3}.

Table C-9 Classifier outputs using two credibility classes

Weighting aggregation model – C4.5 decision tree

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

topic <= 2
|   content_wordsno <= 9
|   |   content_charno <= 89: 1 (4.0)
|   |   content_charno > 89: 3 (4.0/1.0)
|   content_wordsno > 9: 1 (50.0/1.0)
topic > 2
|   author_verif <= 0
|   |   author_descrelate <= 0
|   |   |   topic <= 7
|   |   |   |   content_hasURL <= 0
```

```
|   |   |   |   |   topic <= 4: 1 (14.0)
|   |   |   |   |   topic > 4
|   |   |   |   |   |   content_formal <= 0
|   |   |   |   |   |   |   content_excmarkno <= 0
|   |   |   |   |   |   |   |   content_hasimage <= 0
|   |   |   |   |   |   |   |   |   author_yearsntwt <= 1: 3 (2.0)
|   |   |   |   |   |   |   |   |   author_yearsntwt > 1: 1 (14.0/1.0)
|   |   |   |   |   |   |   |   content_hasimage > 0
|   |   |   |   |   |   |   |   |   author_news <= 0: 3 (6.0)
|   |   |   |   |   |   |   |   |   author_news > 0: 1 (4.0/1.0)
|   |   |   |   |   |   |   content_excmarkno > 0: 1 (3.0)
|   |   |   |   |   |   content_formal > 0: 3 (2.0)
|   |   |   |   |   content_hasURL > 0
|   |   |   |   |   |   author_hasbio <= 0: 3 (2.0)
|   |   |   |   |   |   author_hasbio > 0
|   |   |   |   |   |   |   content_delicate <= 0
|   |   |   |   |   |   |   |   content_wordsno <= 8: 1 (3.0)
|   |   |   |   |   |   |   |   content_wordsno > 8
|   |   |   |   |   |   |   |   |   content_rank <= 4: 1 (3.0/1.0)
|   |   |   |   |   |   |   |   |   content_rank > 4: 3 (8.0)
|   |   |   |   |   |   |   content_delicate > 0: 1 (3.0)
|   |   |   |   topic > 7: 1 (8.0)
|   |   |   author_descrelate > 0
|   |   |   |   content_emotino <= 0
|   |   |   |   |   author_emp <= 0
|   |   |   |   |   |   content_favno <= 0: 3 (8.0)
|   |   |   |   |   |   content_favno > 0
|   |   |   |   |   |   |   content_rank <= 2: 3 (3.0/1.0)
|   |   |   |   |   |   |   content_rank > 2: 1 (5.0)
|   |   |   |   |   author_emp > 0: 3 (7.0)
|   |   |   |   content_emotino > 0: 1 (2.0)
|   |   author_verif > 0
|   |   |   content_charno <= 97: 1 (2.0)
|   |   |   content_charno > 97: 3 (7.0)

Number of Leaves  :     23

Size of the tree :      45


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         110               67.0732 %
Incorrectly Classified Instances        54               32.9268 %
Kappa statistic                          0.2398
Mean absolute error                      0.3457
Root mean squared error                  0.5518
Relative absolute error                 80.4864 %
Root relative squared error            119.1857 %
Coverage of cases (0.95 level)          76.8293 %
Mean rel. region size (0.95 level)      67.0732 %
Total Number of Instances              164

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area  PRC Area  Class
                 0.752     0.510     0.766       0.752    0.759       0.240    0.586     0.705     1
                 0.490     0.248     0.472       0.490    0.481       0.240    0.586     0.425     3
Weighted Avg.    0.671     0.428     0.674       0.671    0.672       0.240    0.586     0.618

=== Confusion Matrix ===

  a  b   <-- classified as
 85 28 |  a = 1
 26 25 |  b = 3
```

Weighting aggregation model – SVM (SMO)

```
=== Classifier model (full training set) ===

SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

Classifier for classes: 1, 3

BinarySMO

Machine linear: showing attribute weights, not support vectors.

        1.2089 * (normalized) topic
 +     -0.289  * (normalized) content_rank
 +      0.163  * (normalized) content_retweetno
 +     -0.8392 * (normalized) content_favno
 +      0.7442 * (normalized) content_hashno
 +     -0.1159 * (normalized) content_spellno
 +      0.9119 * (normalized) content_qmarkno
 +     -0.8353 * (normalized) content_excmarkno
 +     -0.0854 * (normalized) content_emotino
 +      0.3249 * (normalized) content_specialno
 +     -0.6838 * (normalized) content_charno
 +      0.2794 * (normalized) content_wordsno
 +      0.5346 * (normalized) content_hasURL
 +      0.9406 * (normalized) content_hasimage
 +     -0.4765 * (normalized) content_pronnoT
 +      0.2717 * (normalized) content_pronounsD
 +     -0.0881 * (normalized) content_squest
```

```
 +      -0.1786 * (normalized) content_haslaugh
 +      -0.2257 * (normalized) content_delicate
 +      -0.0732 * (normalized) content_bad_swear
 +       0.6127 * (normalized) content_religious
 +       0.1578 * (normalized) content_alldelicate
 +      -0.1338 * (normalized) content_hasurgnews
 +       0.1834 * (normalized) content_formal
 +       1.5043 * (normalized) author_verif
 +      -1.2668 * (normalized) author_fwngno
 +       0.4028 * (normalized) author_flrno
 +      -0.3768 * (normalized) author_logflrno
 +      -0.2861 * (normalized) author_ratiofwfl
 +      -0.0636 * (normalized) author_tweetsno
 +       0.2149 * (normalized) author_favno
 +      -0.8904 * (normalized) author_ratiofavtweet
 +      -0.1798 * (normalized) author_news
 +       0.0922 * (normalized) author_hasbio
 +      -0.5571 * (normalized) author_edu
 +       0.2137 * (normalized) author_emp
 +       0.2912 * (normalized) author_contact
 +      -0.0261 * (normalized) author_allinf1
 +      -0.2965 * (normalized) author_allinf2
 +       0.3635 * (normalized) author_hasweb
 +       0.4558 * (normalized) author_yearsntwt
 +       0.9909 * (normalized) author_descrelate
 +       0.4525 * (normalized) author_locrelate
 +      -0.3483 * (normalized) author_allrelate
 +       0.0274 * (normalized) author_specialch
 -       1.7749

Number of kernel evaluations: 9486 (91.595% cached)



Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           122               74.3902 %
Incorrectly Classified Instances         42                25.6098 %
Kappa statistic                          0.3537
Mean absolute error                      0.2561
Root mean squared error                  0.5061
Relative absolute error                  59.6186 %
Root relative squared error              109.3073 %
Coverage of cases (0.95 level)           74.3902 %
Mean rel. region size (0.95 level)       50      %
Total Number of Instances                164

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area  PRC Area  Class
                 0.876     0.549     0.780       0.876    0.825       0.362    0.664     0.768     1
                 0.451     0.124     0.622       0.451    0.523       0.362    0.664     0.451     3
Weighted Avg.    0.744     0.417     0.730       0.744    0.731       0.362    0.664     0.670

=== Confusion Matrix ===

  a  b   <-- classified as
 99 14 |  a = 1
 28 23 |  b = 3
```

Weighting aggregation model – Random Forest tree - Relief Algorithm feature set

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomForest -I 100 -K 0 -S 1 -num-slots 1
Relation:                              CRED-weka.filters.supervised.attribute.AttributeSelection-
Eweka.attributeSelection.ReliefFAttributeEval -M -1 -D 1 -K 10-Sweka.attributeSelection.Ranker
-T -1.7976931348623157E308 -N -1-weka.filters.unsupervised.attribute.Remove-R13-46
Instances:    164
Attributes:   13
              content_hasimage
              topic
              content_hasURL
              author_locrelate
              author_descrelate
              author_emp
              content_alldelicate
              content_wordsno
              author_hasweb
              author_logflrno
              author_verif
              content_hasurgnews
              sum
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 100 trees, each constructed while considering 4 random features.
Out of bag error: 0.2561



Time taken to build model: 0.29 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           127               77.439  %
```

```
Incorrectly Classified Instances         37               22.561  %
Kappa statistic                          0.4404
Mean absolute error                      0.3184
Root mean squared error                  0.4055
Relative absolute error                 74.1258 %
Root relative squared error             87.5846 %
Coverage of cases (0.95 level)          99.3902 %
Mean rel. region size (0.95 level)      91.7683 %
Total Number of Instances              164

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC       ROC Area  PRC Area
Class
                 0.885     0.471     0.806       0.885    0.844       0.447     0.804     0.901
1
                 0.529     0.115     0.675       0.529    0.593       0.447     0.804     0.625
3
Weighted Avg.    0.774     0.360     0.766       0.774    0.766       0.447     0.804     0.815

=== Confusion Matrix ===

   a   b   <-- classified as
 100  13 |   a = 1
  24  27 |   b = 3
```

- **Effect of majority voting level on classification Accuracy**: A complete classifier outputs of the classification results using labelling obtained by (labellers and experts) for all percentage of majority voting options is listed below in Table C-10

Table C-10 Classifier outputs using all majority voting ratio levels

High percentage of majority voting class (>50%): 66.6667%

96 Instances: 1:70, 2:9, 3:17

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

author_verif <= 0
|   content_hasimage <= 0
|   |   content_charno <= 63
|   |   |   author_descrelate <= 0
|   |   |   |   content_charno <= 58: 1 (2.0)
|   |   |   |   content_charno > 58: 3 (2.0)
|   |   |   author_descrelate > 0: 2 (2.0)
|   |   content_charno > 63: 1 (64.0/7.0)
|   content_hasimage > 0
|   |   author_yearsntwt <= 1: 2 (2.0)
|   |   author_yearsntwt > 1
|   |   |   author_favno <= 831: 1 (9.0/1.0)
|   |   |   author_favno > 831: 3 (5.0/1.0)
author_verif > 0
|   content_qmarkno <= 0: 3 (8.0/3.0)
|   content_qmarkno > 0: 2 (2.0)

Number of Leaves  :     9

Size of the tree :      17


Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          64               66.6667 %
Incorrectly Classified Instances        32               33.3333 %
Kappa statistic                          0.1198
Mean absolute error                      0.2527
Root mean squared error                  0.4449
Relative absolute error                 86.8631 %
Root relative squared error            117.6166 %
Coverage of cases (0.95 level)          77.0833 %
Mean rel. region size (0.95 level)      55.2083 %
Total Number of Instances               96

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC       ROC Area  PRC Area
Class
                 0.871     0.654     0.782       0.871    0.824       0.248     0.620     0.760
1
                 0.111     0.046     0.200       0.111    0.143       0.085     0.413     0.118
2
                 0.118     0.139     0.154       0.118    0.133       -0.024    0.395     0.163
3
Weighted Avg.    0.667     0.506     0.616       0.667    0.638       0.184     0.561     0.594
```

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 61  2  7 |  a = 1
  4  1  4 |  b = 2
 13  2  2 |  c = 3
```

Low percentage of majority voting class (<=50%) - Maj_Class2: 45.6311%

103 data Instances - 1:37, 2:44, 3:22

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

topic <= 2
|   content_pronounsD <= 0
|   |   content_alldelicate <= 0
|   |   |   content_hasURL <= 0: 2 (3.0)
|   |   |   content_hasURL > 0
|   |   |   |   content_charno <= 87: 2 (2.0)
|   |   |   |   content_charno > 87
|   |   |   |   |   content_rank <= 14: 1 (5.0)
|   |   |   |   |   content_rank > 14
|   |   |   |   |   |   content_retweetno <= 1: 1 (2.0)
|   |   |   |   |   |   content_retweetno > 1: 2 (2.0)
|   |   content_alldelicate > 0: 1 (2.0)
|   content_pronounsD > 0: 1 (6.0)
topic > 2
|   author_verif <= 0
|   |   content_hasurgnews <= 0
|   |   |   author_favno <= 167
|   |   |   |   content_emotino <= 0
|   |   |   |   |   content_delicate <= 0
|   |   |   |   |   |   content_pronounsD <= 0
|   |   |   |   |   |   |   author_descrelate <= 0
|   |   |   |   |   |   |   |   content_formal <= 0
|   |   |   |   |   |   |   |   |   author_yearsntwt <= 0: 2 (2.0)
|   |   |   |   |   |   |   |   |   author_yearsntwt > 0
|   |   |   |   |   |   |   |   |   |   author_hasweb <= 0
|   |   |   |   |   |   |   |   |   |   |   author_flrno <= 5994: 1 (7.0)
|   |   |   |   |   |   |   |   |   |   |   author_flrno > 5994: 3 (2.0)
|   |   |   |   |   |   |   |   |   |   author_hasweb > 0
|   |   |   |   |   |   |   |   |   |   |   content_charno <= 129
|   |   |   |   |   |   |   |   |   |   |   |   topic <= 4: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   topic > 4: 2 (3.0)
|   |   |   |   |   |   |   |   |   |   |   content_charno > 129: 3 (2.0)
|   |   |   |   |   |   |   |   content_formal > 0: 2 (2.0)
|   |   |   |   |   |   |   author_descrelate > 0: 2 (6.0/1.0)
|   |   |   |   |   |   content_pronounsD > 0
|   |   |   |   |   |   |   author_yearsntwt <= 2: 1 (2.0)
|   |   |   |   |   |   |   author_yearsntwt > 2: 3 (4.0)
|   |   |   |   |   content_delicate > 0
|   |   |   |   |   |   content_hasimage <= 0: 1 (7.0)
|   |   |   |   |   |   content_hasimage > 0: 3 (3.0/1.0)
|   |   |   |   content_emotino > 0: 1 (2.0)
|   |   |   author_favno > 167
|   |   |   |   author_favno <= 562
|   |   |   |   |   author_hasweb <= 0: 3 (5.0/1.0)
|   |   |   |   |   author_hasweb > 0
|   |   |   |   |   |   content_hashno <= 1: 3 (3.0/1.0)
|   |   |   |   |   |   content_hashno > 1: 2 (9.0/2.0)
|   |   |   author_favno > 562: 2 (7.0)
|   |   content_hasurgnews > 0
|   |   |   author_favno <= 21: 3 (2.0)
|   |   |   author_favno > 21: 1 (2.0)
|   author_verif > 0: 2 (9.0/1.0)

Number of Leaves  :     27

Size of the tree :      53


Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          47               45.6311 %
Incorrectly Classified Instances        56               54.3689 %
Kappa statistic                          0.1423
Mean absolute error                      0.377
Root mean squared error                  0.5654
Relative absolute error                 87.7907 %
Root relative squared error            122.0274 %
Coverage of cases (0.95 level)          65.0485 %
Mean rel. region size (0.95 level)      54.3689 %
Total Number of Instances              103

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area  PRC Area  Class
                 0.459     0.288     0.472       0.459     0.466       0.173     0.624     0.425     1
                 0.591     0.390     0.531       0.591     0.559       0.199     0.575     0.463     2
                 0.182     0.173     0.222       0.182     0.200       0.010     0.532     0.241     3
Weighted Avg.    0.456    0.307    0.444        0.456     0.449       0.149     0.583     0.402
```

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 17 12  8 |  a = 1
 12 26  6 |  b = 2
  7 11  4 |  c = 3
```

Low percentage of majority voting class (<=50%) - Maj_Low: 46.6019%

103 Instances- 1:58:2:23,3:22

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

author_verif <= 0
|   topic <= 2: 1 (20.0/1.0)
|   topic > 2
|   |   content_excmarkno <= 0
|   |   |   author_favno <= 237
|   |   |   |   content_spellno <= 0
|   |   |   |   |   author_descrelate <= 0
|   |   |   |   |   |   content_hasurgnews <= 0
|   |   |   |   |   |   |   content_hasimage <= 0
|   |   |   |   |   |   |   |   author_tweetsno <= 60415: 1 (17.0/1.0)
|   |   |   |   |   |   |   |   author_tweetsno > 60415: 3 (4.0/1.0)
|   |   |   |   |   |   |   content_hasimage > 0
|   |   |   |   |   |   |   |   author_ratiofavtweet <= 106.630873: 1 (3.0)
|   |   |   |   |   |   |   |   author_ratiofavtweet > 106.630873: 3 (4.0)
|   |   |   |   |   |   content_hasurgnews > 0: 3 (2.0)
|   |   |   |   |   author_descrelate > 0
|   |   |   |   |   |   content_wordsno <= 14: 2 (3.0)
|   |   |   |   |   |   content_wordsno > 14
|   |   |   |   |   |   |   content_hasimage <= 0: 1 (4.0)
|   |   |   |   |   |   |   content_hasimage > 0: 3 (4.0)
|   |   |   |   content_spellno > 0: 1 (6.0)
|   |   |   author_favno > 237
|   |   |   |   content_pronnoT <= 0
|   |   |   |   |   content_hasimage <= 0
|   |   |   |   |   |   author_hasweb <= 0: 2 (4.0)
|   |   |   |   |   |   author_hasweb > 0
|   |   |   |   |   |   |   author_fwngno <= 54: 2 (2.0)
|   |   |   |   |   |   |   author_fwngno > 54
|   |   |   |   |   |   |   |   content_wordsno <= 15: 3 (3.0)
|   |   |   |   |   |   |   |   content_wordsno > 15: 2 (4.0/1.0)
|   |   |   |   |   content_hasimage > 0
|   |   |   |   |   |   author_yearsntwt <= 2: 1 (3.0/1.0)
|   |   |   |   |   |   author_yearsntwt > 2: 3 (2.0)
|   |   |   |   content_pronnoT > 0: 1 (3.0/1.0)
|   |   content_excmarkno > 0
|   |   |   content_charno <= 118: 2 (2.0)
|   |   |   content_charno > 118: 1 (2.0)
author_verif > 0
|   content_pronnoT <= 0: 2 (9.0/2.0)
|   content_pronnoT > 0: 1 (2.0)

Number of Leaves  :     21

Size of the tree :      41


Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          48               46.6019 %
Incorrectly Classified Instances        55               53.3981 %
Kappa statistic                          0.0854
Mean absolute error                      0.3725
Root mean squared error                  0.5629
Relative absolute error                 94.6404 %
Root relative squared error            127.0664 %
Coverage of cases (0.95 level)          65.0485 %
Mean rel. region size (0.95 level)      55.6634 %
Total Number of Instances              103

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area  PRC Area  Class
                 0.621     0.511     0.610       0.621    0.615       0.110    0.518     0.555     1
                 0.478     0.175     0.440       0.478    0.458       0.295    0.611     0.308     2
                 0.045     0.222     0.053       0.045    0.049       -0.187   0.426     0.183     3
Weighted Avg.    0.466     0.374     0.453       0.466    0.459       0.088    0.519     0.420

=== Confusion Matrix ===

  a  b  c   <-- classified as
 36  7 15 |  a = 1
  9 11  3 |  b = 2
 14  7  1 |  c = 3
```

Low percentage of majority voting class (<=50%) - Maj_Hi: 32.0388%

103 Instances - 1:37, 2:27, 3:39

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

author_verif <= 0
|   author_favno <= 282
|   |   author_tweetsno <= 294: 2 (5.0/1.0)
|   |   author_tweetsno > 294
|   |   |   topic <= 7
|   |   |   |   content_hasimage <= 0
|   |   |   |   |   content_alldelicate <= 0
|   |   |   |   |   |   content_emotino <= 0
|   |   |   |   |   |   |   author_hasbio <= 0: 1 (2.0)
|   |   |   |   |   |   |   author_hasbio > 0
|   |   |   |   |   |   |   |   content_spellno <= 0
|   |   |   |   |   |   |   |   |   author_ratiofwfl <= 0.64
|   |   |   |   |   |   |   |   |   |   author_descrelate <= 0
|   |   |   |   |   |   |   |   |   |   |   author_fwngno <= 10
|   |   |   |   |   |   |   |   |   |   |   |   author_favno <= 0: 3 (4.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   author_favno > 0: 1 (3.0)
|   |   |   |   |   |   |   |   |   |   |   author_fwngno > 10: 3 (7.0)
|   |   |   |   |   |   |   |   |   |   author_descrelate > 0
|   |   |   |   |   |   |   |   |   |   |   content_rank <= 10: 1 (3.0)
|   |   |   |   |   |   |   |   |   |   |   content_rank > 10: 3 (2.0)
|   |   |   |   |   |   |   |   |   author_ratiofwfl > 0.64: 1 (5.0)
|   |   |   |   |   |   |   |   content_spellno > 0: 1 (3.0)
|   |   |   |   |   |   content_emotino > 0: 1 (2.0)
|   |   |   |   |   content_alldelicate > 0: 1 (8.0)
|   |   |   |   content_hasimage > 0
|   |   |   |   |   author_flrno <= 471: 1 (2.0)
|   |   |   |   |   author_flrno > 471: 3 (12.0/1.0)
|   |   |   topic > 7
|   |   |   |   author_hasweb <= 0
|   |   |   |   |   content_charno <= 113: 3 (2.0)
|   |   |   |   |   content_charno > 113: 1 (2.0)
|   |   |   |   author_hasweb > 0: 2 (4.0)
|   author_favno > 282
|   |   author_emp <= 0
|   |   |   content_hasimage <= 0
|   |   |   |   author_allrelate <= 0: 2 (4.0)
|   |   |   |   author_allrelate > 0
|   |   |   |   |   content_alldelicate <= 0
|   |   |   |   |   |   content_hashno <= 2: 3 (3.0)
|   |   |   |   |   |   content_hashno > 2: 2 (8.0/1.0)
|   |   |   |   |   content_alldelicate > 0: 3 (2.0)
|   |   |   content_hasimage > 0: 3 (6.0/3.0)
|   |   author_emp > 0: 3 (3.0/1.0)
author_verif > 0
|   content_hashno <= 1: 3 (4.0/2.0)
|   content_hashno > 1: 2 (7.0/1.0)

Number of Leaves  :     24

Size of the tree :      47


Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          33               32.0388 %
Incorrectly Classified Instances        70               67.9612 %
Kappa statistic                         -0.038
Mean absolute error                      0.4326
Root mean squared error                  0.6047
Relative absolute error                 98.382  %
Root relative squared error            128.9324 %
Coverage of cases (0.95 level)          64.0777 %
Mean rel. region size (0.95 level)      56.6343 %
Total Number of Instances              103

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC       ROC Area  PRC Area  Class
                 0.351     0.394     0.333       0.351     0.342       -0.042    0.564     0.398     1
                 0.222     0.224     0.261       0.222     0.240       -0.002    0.542     0.276     2
                 0.359     0.422     0.341       0.359     0.350       -0.062    0.489     0.367     3
Weighted Avg.    0.320    0.360     0.317       0.320     0.318        -0.039   0.530     0.354

=== Confusion Matrix ===

  a  b  c   <-- classified as
 13  9 15 |  a = 1
  9  6 12 |  b = 2
 17  8 14 |  c = 3
```

Low percentage of majority voting class (<=50%) - Maj_N: 41.7476%

103 Instances - 1:38, 2:43, 3:22

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

topic <= 2
|   content_pronounsD <= 0
|   |   content_alldelicate <= 0
|   |   |   content_hasURL <= 0: 2 (3.0)
|   |   |   content_hasURL > 0
|   |   |   |   content_charno <= 87: 2 (2.0)
|   |   |   |   content_charno > 87
|   |   |   |   |   content_rank <= 14: 1 (5.0)
|   |   |   |   |   content_rank > 14
|   |   |   |   |   |   content_retweetno <= 1: 1 (2.0)
|   |   |   |   |   |   content_retweetno > 1: 2 (2.0)
|   |   content_alldelicate > 0: 1 (2.0)
|   content_pronounsD > 0: 1 (6.0)
topic > 2
|   author_verif <= 0
|   |   content_hasurgnews <= 0
|   |   |   author_favno <= 167
|   |   |   |   content_emotino <= 0
|   |   |   |   |   content_delicate <= 0
|   |   |   |   |   |   content_pronounsD <= 0
|   |   |   |   |   |   |   author_descrelate <= 0
|   |   |   |   |   |   |   |   author_hasweb <= 0
|   |   |   |   |   |   |   |   |   author_flrno <= 7030: 1 (8.0)
|   |   |   |   |   |   |   |   |   author_flrno > 7030: 3 (3.0/1.0)
|   |   |   |   |   |   |   |   author_hasweb > 0
|   |   |   |   |   |   |   |   |   content_charno <= 129
|   |   |   |   |   |   |   |   |   |   topic <= 4: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   topic > 4: 2 (5.0)
|   |   |   |   |   |   |   |   |   content_charno > 129: 3 (2.0)
|   |   |   |   |   |   |   author_descrelate > 0: 2 (6.0/1.0)
|   |   |   |   |   |   content_pronounsD > 0
|   |   |   |   |   |   |   author_yearsntwt <= 2: 1 (2.0)
|   |   |   |   |   |   |   author_yearsntwt > 2: 3 (4.0)
|   |   |   |   |   content_delicate > 0
|   |   |   |   |   |   content_hasimage <= 0: 1 (7.0)
|   |   |   |   |   |   content_hasimage > 0: 3 (3.0/1.0)
|   |   |   |   content_emotino > 0: 1 (2.0)
|   |   |   author_favno > 167
|   |   |   |   author_favno <= 562
|   |   |   |   |   author_hasweb <= 0: 3 (5.0/1.0)
|   |   |   |   |   author_hasweb > 0
|   |   |   |   |   |   content_hashno <= 1: 3 (3.0/1.0)
|   |   |   |   |   |   content_hashno > 1: 2 (9.0/2.0)
|   |   |   |   author_favno > 562: 2 (7.0)
|   |   content_hasurgnews > 0
|   |   |   author_favno <= 21: 3 (2.0)
|   |   |   author_favno > 21: 1 (2.0)
|   author_verif > 0: 2 (9.0/1.0)

Number of Leaves  :     25

Size of the tree :      49


Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          43               41.7476 %
Incorrectly Classified Instances        60               58.2524 %
Kappa statistic                          0.0892
Mean absolute error                      0.3844
Root mean squared error                  0.5707
Relative absolute error                 89.3798 %
Root relative squared error            123.0709 %
Coverage of cases (0.95 level)          62.1359 %
Mean rel. region size (0.95 level)      55.9871 %
Total Number of Instances              103

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area  PRC Area  Class
                 0.395     0.338     0.405       0.395     0.400       0.057     0.538     0.429     1
                 0.581     0.350     0.543       0.581     0.562       0.230     0.621     0.496     2
                 0.136     0.210     0.150       0.136     0.143      -0.076     0.529     0.241     3
Weighted Avg.    0.417     0.316     0.408       0.417     0.413       0.100     0.571     0.417

=== Confusion Matrix ===

  a  b  c   <-- classified as
 15 12 11 |  a = 1
 12 25  6 |  b = 2
 10  9  3 |  c = 3
```

Appendix D. **List of Published Papers**

The material presented in this thesis is largely based on the on the following peer-reviewed published papers, which are referred to throughout the thesis:

Chapter 1 of the thesis was inspired by the listed paper below. This is a joint work between both authors.

1. Madlberger, Lisa, and Amal Almansour. "Predictions based on Twitter—A critical view on the research process." In Data and Software Engineering (ICODSE), 2014 International Conference on, pp. 1-6. IEEE, 2014.

Chapter 2 of the thesis which covers the review of literature is largely based on the listed paper. Most parts of the paper have been written by the main author with significant contribution from the second author.

2. AlMansour, Amal Abdullah, Ljiljana Brankovic, and Costas S. Iliopoulos. "Evaluation of credibility assessment for microblogging: models and future directions." Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business. ACM, 2014.

Selective parts from chapters 1, 2, and 5 of the thesis are based on the following papers:

3. AlMansour, Amal Abdullah, Ljiljana Brankovic, and Costas S. Iliopoulos. "A Model for Recalibrating Credibility in Different Contexts and Languages-A Twitter Case Study." International Journal of Digital Information and Wireless Communications (IJDIWC) 4, no. 1 (2014): 53-62.
4. AlMansour, Amal Abdullah. "Towards Customizing Credibility in Different Contexts: Languages, Topics and Locations-A Twitter Case Study." The International Conference on Digital Information Processing, E-Business and Cloud Computing (DIPECC2013). The Society of Digital Information and Wireless Communication, 2013.

Chapter 5 of the thesis is based on the paper below, and most parts of the paper have been written by the main author including all data analysis and relevant discussions.

5. AlMansour, Amal Abdullah, and Costas S. Iliopoulos. "Using Arabic Microblogs Features in Determining Credibility." In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 1212-1219. ACM, 2015.

Appendix E. **Ethical Approval**

The ethical approval required for this research study is obtained after meeting all the requirements of the BDM RESC. A letter granting full ethical approval (BDM/14/15-9) for the study is listed in below.

Amal Abdullah Almansour
744
4 St George Wharf
Vauxhall
SW8 2JF

19 November 2014

Dear Amal Abdullah Almansour

BDM/14/15-9 Credibility Assessment for Arabic Social Media

Review Outcome: Full Approval

Thank you for sending in the amendments/clarifications requested to the above project. I am pleased to inform you that these meet the requirements of the BDM RESC and therefore that full approval is now granted.

Please ensure that you follow all relevant guidance as laid out in the King's College London Guidelines on Good Practice in Academic Research (http://www.kcl.ac.uk/college/policyzone/index.php?id=247).

For your information ethical approval is granted until 19/11/2016. If you need approval beyond this point you will need to apply for an extension to approval at least two weeks prior to this explaining why the extension is needed, (please note however that a full re-application will not be necessary unless the protocol has changed). You should also note that if your approval is for one year, you will not be sent a reminder when it is due to lapse.

Ethical approval is required to cover the duration of the research study, up to the conclusion of the research. The conclusion of the research is defined as the final date or event detailed in the study description section of your approved application form (usually the end of data collection when all work with human participants will have been completed), not the completion of data analysis or publication of the results.

For projects that only involve the further analysis of pre-existing data, approval must cover any period during which the researcher will be accessing or evaluating individual sensitive and/or un-anonymised records.

Note that after the point at which ethical approval for your study is no longer required due to the study being complete (as per the above definitions), you will still need to ensure all research data/records management and storage procedures agreed to as part of your application are adhered to and carried out accordingly.

If you do not start the project within three months of this letter please contact the Research Ethics Office.

Should you wish to make a modification to the project or request an extension to approval you will need approval for this and should follow the guidance relating to modifying approved applications:
http://www.kcl.ac.uk/innovation/research/support/ethics/applications/modifications.aspx

Please would you also note that we may, for the purposes of audit, contact you from time to time to ascertain the status of your research.

If you have any query about any aspect of this ethical approval, please contact your panel/committee administrator in the first instance (http://www.kcl.ac.uk/innovation/research/support/ethics/contact.aspx)

We wish you every success with this work.

Yours sincerely,

Tom Billins, Senior Research Ethics Officer
For and on behalf of
Dr Blánaid Daly, Chair
Biomedical Sciences, Dentistry, Medicine and Natural and Mathematical Sciences Research Ethics Subcommittee (BDM RESC)
Cc. Costas Iliopoulos
